Naar een rationeel systeem voor toetsing van studieprestaties in probleemgestuurd medisch onderwijs : studies naar betrouwbaarheid en validiteit van toetsen voor praktische vaardigheden

NAAR EEN RATIONEEL SYSTEEM VOOR TOETSING VAN STUDIEPRESTATIES IN PROBLEEMGESTUURD MEDISCH ONDERWIJS

Studies naar betrouwbaarheid en validiteit
van toetsen voor praktische vaardigheden

NAAR EEN RATIONEEL SYSTEEM VOOR TOETSING VAN STUDIEPRESTATIES IN PROBLEEMGESTUURD MEDISCH ONDERWIJS

Studies naar betrouwbaarheid en validiteit van toetsen voor praktische vaardigheden

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de Rector Magnificus, Prof. Dr. F.I.M. Bonke,
volgens het besluit van het College van Dekanen,
in het openbaar te verdedigen op vrijdag,
15 september 1989 om 14.00 uur

door

Cornelis Petronella Maria van der Vleuten

geboren te Eindhoven in 1956

Promotor:	Prof. dr. W.H.F.W. Wijnen
Beoordelingscommissie:	Prof. dr. C.P.A. van Boven (voorzitter) Prof. dr. ir. A. Hasman Prof. dr. G.J. Mellenbergh Prof. dr. J. Moll Prof. dr. H.G. Schmidt

Voor Marianne, Susan, Lotte & Maaike

Voor mijn ouders

Inhoudsopgave

Voorwoord

Dit proefschrift is in meerdere opzichten een typisch Maastrichts produkt. In de eerste plaats vanwege de inhoud. Het proefschrift handelt over probleemgestuurd onderwijs, meer specifiek over de wijze waarop de evaluatie van studieprestaties in dit onderwijssysteem is aangepakt. Wellicht meer nog dan de algemene onderwijsmethodiek wordt de evaluatie van studieprestaties gekenmerkt door een unieke Maastrichtse invulling. Hopelijk proeft de lezer in deze publicatie dat de evaluatie-aanpak, c.q. de achterliggende gedachtes en rationales, van groot belang zijn voor het uiteindelijk te realiseren toetssysteem, waarschijnlijk van groter belang dan de actuele inhoud van dit proefschrift of de wetenschappelijke output ervan.

In de tweede plaats vanwege de wijze waarop dit proefschrift tot stand is gekomen. De Maastrichtse onderwijs- (en onderzoeks-) organisatie kent via haar matrix-management systeem interdisciplinaire projecten. In het onderwijs (en onderzoek) te verrichten taken zijn zelden mono-disciplinair, maar vergen teamwerk en de deskundigheid van meerdere disciplines. Dit proefschrift is een blijk van de wijze waarop een dergelijke samenwerking vruchtbaar kan zijn, van de wijze waarop disciplinegrenzen zoals geneeskunde, onderwijskunde en psychometrie vervagen wanneer aan een gezamenlijke taak wordt gewerkt. Grote dank ben ik dan ook verschuldigd aan alle betrokkenen van het Project Evaluatie van Studieresultaten. De bovengenoemde aanpak, dit proefschrift, mijn over de jaren verworven deskundigheid, het bestaande evaluatiesysteem, de huidige wetenschappelijke productie, het plezier in het werk, zij zijn allen voortgekomen uit het project. Wynand, Maarten, Tjaart, Hetty, Betsy, Scheltus, Paulien, Mirjam, Marian, Yvon, Noël, Math, Carla, Monique, Marjan, Nellie, Nico, Trees, Ron, Jef, Marianne, Pieter, Carla, Henny, Annemarie, Tim, Harry, Daniël, Leo allen die over de jaren betrokken zijn geweest, bedankt.

Naar enkele mensen in het bijzonder gaat een speciale dank uit. Aan Wynand Wijnen, mijn promotor, die door zijn creatieve, 'laterale' denkvermogen een grote inspiratiebron is, genoeg voor een levenswerk aan onderwijsontwikkeling en -onderzoek; aan Scheltus van Luyk, mijn onderzoeks-sparringpartner, in de overtuiging dat binnen afzienbare tijd een soortgelijk document voorhanden komt van zijn hand; aan Maarten Verwijnen, wiens onderwijs-genius mijn ontwikkeling heeft bepaald en aan Ron Hoogenboom, wiens accuratesse op het gebied van (grootschalige) data-analyse de mijne uitstekend compenseerde. Speciale dank ook aan Dave Swanson, voor mij de absolute meester in deskundigheid op het gebied van 'assessment in medical education', en wiens sabbatical leave in Maastricht een voor mij persoonlijk zeer belangrijke leerperiode is geweest. Petry Thiemann wil ik danken voor de voortreffelijke wijze waarop ze voor de opmaak van dit proefschrift heeft gezorgd. De vakgroep O&O ben ik erkentelijk voor de ruimte die zij mij in de afgelopen jaren gegeven heeft.

Tenslotte voor mijn vrouw: Marianne, je altijd kritische commentaar op de inhoud (en taal) van mijn stukken werd zeer gewaardeerd, ook al bleek dat vaak niet op elk moment. Of de drukte na een proefschrift nu echt voorbij zal zijn, daar ben ik wat pessimistisch over. Wanneer ik echter in staat ben evenveel support te geven aan jouw werk als jij aan het mijne, dan ben ik voor de toekomst gerust. Susan, Lotte en Maaike zorgen voor de rest.

Inleiding

Toen in 1974 in Maastricht gestart werd met een probleem-gestuurd medisch curriculum, was nog nauwelijks vastgesteld op welke wijze studieprestaties in dit nieuwe systeem zouden moeten worden geëvalueerd. De Canadese McMaster universiteit, waarop het onderwijssysteem grotendeels gebaseerd was, bood in dit opzicht niet veel houvast. Weliswaar kende McMaster eigen evaluatieprocedures, maar de kern van de toetsing bestond uit de deelname aan de nationale examens van Canada, waardoor de behoefte of noodzaak voor de ontwikkeling van een integraal toetsingssysteem minder sterk aanwezig was. Maastricht stond hier dus voor een nieuwe taak, omdat nationale examens in Nederland nu eenmaal ontbreken.

Het belang van aangepaste toetsing voor het welslagen van het probleemgestuurde onderwijs werd met elk jaar na de start duidelijker. Er was begonnen met een betrekkelijk conventionele benadering, bestaande uit de introductie een stelsel van toetsen als afsluiting van iedere onderwijsleerperiode. Daardoor ontstonden echter spanningen tussen de intenties van het onderwijssysteem en het leergedrag van studenten. Het nieuwe onderwijssysteem streefde een aantal onderwijskundige effecten na, die beter zouden aansluiten bij moderne inzichten over het leren van de student (cf. Schmidt, 1982). Toetsen en examens zijn echter de middelen waarlangs studiesucces kan worden geboekt en dientengevolge oefenen zij een zeer dwingende invloed uit op het leergedrag van studenten (Newble & Jaeger, 1983; Newble & Entwistle, 1986; Frederiksen, 1984). Een conventionele, niet op het onderwijssysteem aansluitend toetsingssysteem, zou het specifieke karakter van het onderwijs uiteindelijk ondergraven. Aldus ontstond een groeiende noodzaak om een eigen aangepast toetsingssysteem te ontwerpen.

Na een aanpak met verscheidene aparte projectgroepen met deeltaken op evaluatiegebied, besloot de Faculteit der Geneeskunde in 1982 alle inzet te bundelen in één multidisciplinair samengestelde groep van artsen en toetsdeskundigen, die de opdracht kreeg de evaluatie van studieprestaties uit te werken, aangepast aan de eisen van het probleem-gestuurde onderwijs (cf. Verwijnen & Imbos, 1982).

Zo ontstond de mogelijkheid om de problematiek van toetsing op een centrale wijze aan te pakken voor een complete faculteit. Deze situatie is met name uniek, omdat in de praktijk van het hoger onderwijs in het algemeen niet of nauwelijks aandacht wordt besteed aan onderwijskundige implicaties van toetsen. Voor zover onderwijsondersteunende instanties binnen onderwijsinstellingen zich met toetsing bezig houden, blijft dit vaak beperkt tot standaard psychometrische toetsanalyses, meestal op grond van vrijwillige deelname van docenten of vakgroepen. Toetsing als onderwijsactiviteit wordt niet zelden opgevat als onderdeel van de vakinhoudelijke competentie van docenten en de

externe bemoeienis ermee wordt soms gezien als een aantasting van deze competentie. Deze geringe aandacht voor onderwijstoetsen heeft niet alleen gevolgen voor de kwaliteit van toetsen en examens, maar leidt in de onderwijspraktijk vaak tot situaties waarin toetsen, in plaats van hulpmiddelen bij de studie, eerder obstakels vormen. Een examensysteem krijgt hiermee vaak het karakter van een hordenloop (De Groot, 1972; Wijnen & Van der Vleuten, 1985).

Door in de facultaire organisatie van het medisch onderwijs te kiezen voor een *gecentraliseerde aanpak* van toetsing ontstond de mogelijkheid de toetsingsproblematiek op een andere wijze te benaderen. Deze aanpak kenmerkt zich voornamelijk door een *rationele benadering*: door kennis te nemen van de bestaande literatuur, door de cumulatie van kennis en eigen ervaringen, door de concrete toepassing van deze kennis op toets- en examenprocedures en door het verrichten van eigen wetenschappelijk onderzoek, wordt deze rationele benadering gefundeerd. Door middel van een interdisciplinair team van 'toetsdeskundigen' worden medisch inhoudelijke, onderwijskundige en psychometrische inzichten zo optimaal mogelijk verenigd. Door een adequate inbedding van deze benadering in de onderwijsorganisatie, wordt de toetsing van studieprestaties een onderdeel van het onderwijs zelf, vormt het een leermiddel voor studenten, kunnen meer adequate beslissingen over studievoortgang worden genomen, en draagt het bij aan de meer algemene kwaliteitsbewaking van het onderwijs (Gijselaers & Van der Vleuten, 1988). De dagelijkse toetspraktijk is daarenboven een bron van gegevens waarmee verder onderzoek kan worden uitgevoerd, en van waaruit verdere onderbouwingen of aanpassingen kunnen worden gemaakt. Het evaluatiesysteem wordt daarmee een *dynamisch systeem*, continu onderhevig aan veranderingen en verbeteringen, op grond van rationele overwegingen.

Deze benadering en de integrale beschrijving van het huidige toetsingssysteem is onderwerp van het eerste hoofdstuk van dit proefschrift. Verslag wordt gedaan van de specifieke eisen die gesteld moeten worden aan een evaluatiesysteem in probleem-gestuurd onderwijs en de wijze waarop dat op dit moment is gerealiseerd. Gepoogd wordt aan te geven dat enkele van de gerealiseerde zaken slechts mogelijk zijn vanwege de centrale benadering van toetsing van studieprestaties.

Dit eerste hoofdstuk vormt de context waarbinnen de overige studies verricht zijn, en waarvan verslag wordt gedaan in de volgende hoofdstukken. Deze hoofdstukken hebben allen betrekking op één aspect van het toetsingssysteem: het gebied van toetsing van praktische vaardigheden. In het algemeen kan gesteld worden dat toetsen voor praktische vaardigheden nog maar nauwelijks zijn ontwikkeld en dat zij nadere psychometrische onderbouwing behoeven. In de hoofdstukken 2, 3, 4 en 5 wordt verslag gedaan van enkele empirische studies naar de betrouwbaarheid en validiteit van deze toetsen.

In hoofdstuk 2 wordt gerapporteerd over de reproduceerbaarheid van de met vaardigheidstoetsen verkregen toetsscores. Door cumulatie van vaardigheidstoetsgegevens over verschillende jaren is een tamelijk grote data set ontstaan. Hierover zijn analyses verricht waarbij gebruik gemaakt werd van de generaliseerbaarheidstheorie (Cronbach et al., 1972). Voor zover het materiaal het

toestond, zijn de potentiële foutenbronnen onderzocht en besproken op hun consequenties voor de betrouwbaarheid van de toetsscores.

Hoofdstuk 3 gaat specifieker in op variabelen die de accuratesse van beoordelaars beïnvloeden. In een experiment is de invloed nagegaan van de mate van deskundigheid van de beoordelaar en het effect van training op de accuratesse van de beoordeling.

Hoofdstuk 4 rapporteert over een experiment waarin de waarde van een schriftelijke vorm voor het meten van praktische vaardigheden centraal staat. Om zowel onderwijskundige motieven, als om redenen van betrouwbaarheid zou een adequate schriftelijke toetsvorm relevante additionele meetinformatie over praktische vaardigheden kunnen opleveren. Het experiment omvatte een afname van een kennistoets over vaardigheden, in de vorm van een voortgangstoets.

In hoofdstuk 5 wordt verslag gedaan van een studie waarin werd nagegaan of het tamelijk dwingende karakter van de gehanteerde beoordelingswijze (gedetailleerde criterialijsten) leidt tot discrepanties met het algemene oordeel van beoordelaars. Eventuele discrepanties kunnen mogelijk het gevolg zijn van beoordelingsaspecten welke niet of in mindere mate door criterialijsten worden gedekt en geven aanwijzingen over de validiteit van de toets.

Tenslotte geeft hoofdstuk 6 een overzicht van de (internationale) stand van zaken met betrekking tot toetsen voor praktische vaardigheden in de geneeskunde. De laatste jaren zijn verschillende studies gepubliceerd naar de meeteigenschappen van soortgelijke toetsen voor praktische vaardigheden. In de vorm van een review worden deze studies in dit hoofdstuk met elkaar vergeleken, en worden gemeenschappelijke psychometrische bevindingen geanalyseerd. De praktische en methodologische consequenties hiervan worden besproken.

De hoofdstukken zijn artikelen die reeds eerder zijn gepubliceerd, dan wel ter beoordeling aan tijdschriften zijn aangeboden (hoofdstuk 5 is een hoofdstuk uit een boek). Daardoor is het onvermijdelijk dat er sprake is van enige herhaling van informatie. Na lezing van hoofdstuk 1 kan de lezer alle passages over de uitleg van het toetsingssysteem overslaan, zonder dat betekenisvolle informatie verloren gaat.

Ter inleiding op de hoofdstukken over praktische vaardigheden zijn nog een tweetal aspecten van belang. In de eerste plaats betreft dat een summiere historische beschrijving van de totstandkoming en het gebruik van dit soort van toetsen in medische opleidingen, en in de tweede plaats enkele woorden over de definitie van praktische vaardigheden.

Nadat in de vijftiger en zestiger jaren als gevolg van de schaalvergroting de multiple choice vraagvorm in het onderwijs op grote schaal geïntroduceerd was, ontstond tegelijkertijd een gevoel van onvrede: de multiple choice vraag (en elke andere gesloten vraagvorm) zou een te beperkt competentiegebied bestrijken, te eenzijdig een beroep doen op herkenning van informatie, en te weinig gericht zijn op toepassingsgerichte kennis en probleemoplossingsvaardigheden (McGuire, 1987). Deze onvrede, mede gevoed door het ontstaan van innova-

tieve onderwijsvormen die deze andere competenties beoogden, hebben aanleiding gegeven tot sterke ontwikkelingen op het gebied van nieuwe instrumenten. In de zeventiger jaren ontstond aldus een groot aanbod van goeddeels schriftelijke instrumenten, waarin op een of andere wijze de werkelijkheid werd nagebootst. De belangrijkste exponent van deze simulatie-instrumenten waren de Patient Management Problems (Rimoldi, 1978; McGuire & Solomon, 1976), waarin op schrift aangeboden patiëntenproblemen, al of niet in een vertakte vorm en vaak voorzien van ingenieuze druktechnieken, moesten worden opgelost. Patient Management Problems hebben lange tijd gediend als een belangrijk onderdeel van de nationale arts-examens in Verenigde Staten en Canada.

De resultaten van het empirisch onderzoek dat volgde op de introductie van deze simulatie-instrumenten was ongunstig: toetsscores bleken weinig betrouwbaar en vooral de incrementele validiteit werd onvoldoende beschouwd. Dit laatste doordat consistent zeer hoge correlaties met scores op multiple choice toetsen werden gevonden (Norcini et al., 1986), waardoor toetsscores verkregen met deze nieuwe instrumenten weinig unieke variantie toevoegden. Het gebruik van dit soort van instrumenten voor de onderwijspraktijk werd dan ook afgeraden (Swanson et al., 1987). Ook de meeste 'national boards' hebben inmiddels deze instrumenten niet meer in hun programma en beperken zich tot multiple choice en soortgelijke vraagvormen.

Ondanks de teleurstellende empirische resultaten, werd door deze ontwikkeling een aanzet gegeven tot een nieuw soort streven: toetsen dienden *representatiever* te zijn voor de (onderwijs-)praktijk. Zij dienden meer te beantwoorden aan de taken die ook in werkelijkheid van studenten worden verwacht. Van onderwijstoetsen in medische opleidingen werd een hogere 'ecologische validiteit' (Brunswik, 1956; De Klerk, 1980) wenselijk geacht.

De ontwikkeling van meer representatieve toetsen werd in de zeventiger jaren in Engeland vervolgd met de introductie van Objective Structured Clinical Examinations (OSCEs; Harden et al., 1975; Harden & Gleeson, 1979), en in de Verenigde Staten met de introductie van simulatiepatiënten in examens (Stillman et al., 1976). Kenmerkend voor deze toetsen is dat kandidaten geconfronteerd worden met praktische situaties, en dat aan de hand van observaties een prestatie wordt uitgedrukt in een waardering. Dat de idee van deze meer representatieve toetsen weerklank vond, moge blijken uit het feit dat zij wereldwijd op talloze plaatsen worden geïntroduceerd, en dat zij een belangrijk nieuw aandachtsgebied vormen voor wetenschappelijke bijeenkomsten over toetsing van klinische competentie (Hart, Harden & Walton, 1986; Hart & Harden, 1987).

Vanuit deze historische situatie, en in de specifieke onderwijskundige context van het probleem-gestuurd leren, is vaardigheidstoetsing ontstaan aan de medische faculteit te Maastricht. Het onderwijssysteem van deze opleiding poogt expliciet theorie en praktijk te integreren. Vroegtijdig worden studenten geconfronteerd met praktijksituaties en leren zij, in samenhang met het meer theoretische curriculum, praktische vaardigheden aan. Dat laatste gebeurt vooral in het speciaal daarvoor ingerichte Skillslab. In de eerste vier jaren van hun studie verwerven studenten in een veilige 'laboratorium-omgeving' de vaardigheden, die van belang zijn voor de klinische beroepsuitoefening. In de

laatste twee klinische jaren van de opleiding worden deze vaardigheden toegepast in de klinische stages. Het lag voor de hand om voor de toetsing van dit uitgebreide vaardigheidsonderwijs een keuze te maken voor een representatieve toetsvorm, waarin de directe observatie van de kandidaten centraal zou staan.

Door hiermee reeds in 1982 van start te gaan en door invoering van deze toetsen in het gehele curriculum voor alle studenten, neemt Maastricht een unieke positie in. Nergens anders wordt op zo uitgebreide schaal met dergelijke toetsen gewerkt.

In de loop der tijd is een relatief grote data set ontstaan waarmee onderzoek mogelijk werd. Door dit in eigen huis verrichte onderzoek, en door soortgelijk onderzoek dat recentelijk op andere plaatsen werd uitgevoerd, is meer bekend geworden over deze nieuwe toetsvorm en de betekenis daarvan voor medisch onderwijs.

In de navolgende hoofdstukken wordt nergens een formele definitie gegeven van de competenties die met de meting van vaardigheden worden beoogd. Er wordt een pragmatisch standpunt ingenomen: de toets meet (idealiter) datgene wat in het onderwijs aan bod is gekomen, c.q. dat wat in het onderwijs belangrijk wordt geacht. De vaardigheden die in het onderwijs worden aangeleerd vallen in vier grotere categorieën uiteen: fysisch diagnostische vaardigheden, therapeutische vaardigheden, laboratorium vaardigheden en sociale vaardigheden. Hieruit blijkt reeds dat bij de toetsing van vaardigheden de aard van de processen waarop een beroep wordt gedaan nogal uiteen loopt, variërend van betrekkelijk eenvoudige psychomotorische vaardigheden (zoals bijvoorbeeld injecteren), tot complexe cognitieve vaardigheden (zoals interpretatie van bevindingen na lichamelijk onderzoek) en affectieve vaardigheden (zoals het voeren van een "slecht nieuws gesprek"). De bovenstaande categorieën zijn weer te verdelen in deelvaardigheden, die gezamenlijk een blauwdruk vormen, op basis waarvan de vaardigheidstoets wordt samengesteld. De blauwdruk bepaalt wat wel en niet in de toets wordt opgenomen, en in zoverre kan worden geproken van een operationele definitie van het te meten concept (Ebel, 1961).

Literatuur

Brunswik, E. (1956) *Perception and Representative Design of Psychological Experiments.* Berkeley: University of California Press.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: Wiley.

Ebel, R. (1961) Must all tests be valid? *American Psychologist, 16,* 640-647.

Frederiksen, N. (1984) The real test bias: influences of testing on teaching and learning. *American Psychologist, 39,* 193-202.

Gijselaers, W.H. & Vleuten, C.P.M. van der (1988) Evaluatie aan de medische faculteit van Maastricht. *Onderwijsverslag 1987 Faculteit der Geneeskunde,* Rijksuniversiteit Limburg.

Groot, A.D. de (1972) *Selectie voor en in het Hoger Onderwijs.* Den Haag: Staatsuitgeverij.

Harden, R., Stevenson, M., Downie, W. & Wilson, G. (1975) Assessment of Clinical Competence using objective structured examinations. *British Medical Journal, 1*, 447-451.

Harden, R.M. & Gleeson, F.A. (1979) ASME Medical Education Booklet No. 8. Assessment of clinical competence using an objective structured clinical examination (OSCE), *Medical Education*.

Hart, I.R., Harden, R.M. & Walton, H.J. (1986) *Newer Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Hart I.R. & Harden, R.M. (1987) *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.

Klerk, L.F.W. de (1980) *Het leren van psychomotorische vaardigheden: Een onderwijspsychologische benadering*. Deventer: Van Loghum Slaterus.

McGuire, C. (1987) Written methods for assessing clinical comptence. In: I.R. Hart & R.M. Harden, (Eds.), *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.

McGuire, C.H. & Solomon, C.M. (1976) *Construction and Use of Written Simulations*. Chicago: The Psychological Corporation.

Newble, D.I. & Jaeger, K. (1983) The effect of assessment and examinations on the learning of medical students. *Medical Education, 17*, 165-171.

Newble, D.I. & Entwistle, N.J. (1986) Learning styles and approaches: implications for medical education. *Medical Education, 20*, 162-165.

Norcini, J.J., Swanson, D.B., Grosso, L.J. & Webster, G.D. (1986) The psychometric characteristics of some common item formats. In: I.R. Hart, R.M. Harden & H.J. Walton, (Eds.), *Newer Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Rimoldi, H.J.A. (1961) The test of diagnostic skills. *Journal of Medical Education, 30*, 72-79.

Schmidt, H.G. (1982) *Activatie van voorkennis, intrinsieke motivatie en de verwerking van tekst: Studies in probleemgestuurd onderwijs*. Apeldoorn: Van Walraven, Academisch Proefschrift.

Stillman, P.L., Sabers, D. & Redfield, D. (1976) The use of paraprofessionals to teach and evaluate interviewing skills in medical students. *Pediatrics, 57*, 769-774.

Swanson, D., Norcini, J., & Grosso, L. (1987) Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*, 220-246.

Verwijnen, G.M. & Imbos, Tj. (1982) *Het evaluatiesysteem van de Fakulteit der Geneeskunde: Voorzieningen en verdere ontwikkelingen*. PES-publ. nr. 1, Intern Rapport Rijksuniversiteit Limburg.

Wijnen, W.H.F.W. & Vleuten, C.P.M. van der (1985) Toetsing: Hordenloop of voortgangscontrole? *Universiteit & Hogeschool, 31*, 270-279.

HOOFDSTUK 1

ASSESSMENT IN PROBLEM-BASED LEARNING: THE CASE OF MAASTRICHT

Summary

Problem-based learning is now acknowledged to be a successful educational method, and it has been adopted in many institutions in higher education. However, for problem-based learning to be successful, the system used for assessment of student achievement must be consistent with the educational principles of problem-based learning. Whereas the literature on the assessment of clinical competence supplies a rich source of information, only a few integrative systems for problem-based learning have been described. This paper reports on the assessment system developed and in use in the Maastricht medical school in the Netherlands. The school adopted a non-departmental centralized system of assessment, and tried to translate the educational premises of problem-based learning into an integrative assessment system. A number of assessment principles were specified therefore, to which the assessment program should be adapted.

Four types of competencies are distinguished in the assessment program: knowledge, skills, problem-solving and attitudes. For each of these competency types the formal and informal assessments used are described. The reasons for the choices that were made are given, our experiences with these assessments are delineated, and some results on reliability and validity are summarized. For knowledge and skills the testing system has an elaborate program of formal assessments, consisting of Block Tests, Progress Tests, and Skills Tests. With the exception of Clinical Ratings, the assessment of problem-solving and attitudes are still restricted to informal evaluations.

Design of an assessment system must include more than independent construction and use of testing instruments. The overall goals and organization of the assessment system are also discussed, along with procedures for quality control of test materials and interrelationships with student counseling and promotion.

It is concluded that the current assessment system is consistent with the principles of problem-based learning and meets many of the school's needs, but that further refinement of the system is desirable. Methods for assessment of problem-solving and attitudes are still needed, though results cannot be expected in the short term, because of the current state of the art in measuring these competencies. A centralized approach to assessment permits use of more elaborate testing methods, without loss of control over quality, a likely problem if a traditional, decentralized approach were used. The centralized approach also allows systematic, scientific development of assessment methods that are open to public scrutiny.

Introduction

Over the last decades numerous instructional alternatives have been introduced in modern medical curricula. The instructional approach of problem-based learning is a predominant example of these modern educational alternatives. Judging by the number of medical schools applying problem-based learning, in whole or in part, it can be concluded that the method is at present well-established. The rationale of the method (Barrows & Tamblyn, 1980; Schmidt, 1983; Schmidt & De Volder, 1984; Barrows, 1985; Kaufman, 1985) and to some extent its effectiveness (Schmidt, Dauphinee & Patel, 1987; Kantrowitz et al., 1987) is well if not thoroughly documented. On the other hand, as to the method(s) of assessing student achievements within the context of problem-based learning, less consensus or documentation exists. Although there is ample literature pertaining to the evaluation of clinical competence or problem solving (e.g. Neufeld & Norman, 1985), an area naturally and conceptually related to the problem-based learning method, very few integrative evaluation systems have been proposed. Among the exceptions are Feletti (1980) and West, Umland & Lucero (1985), but generally reference is restricted to isolated specific instruments used in a problem-based learning context (e.g. Powles et al., 1981; Feletti & Engel, 1980; Williams et al., 1983; Barrows, 1985).

Probably, this absence in part can be attributed to the presence of national examination systems in most countries, through which the (external) pressure for an internal assessment system is not severely compelling. Conversely however, these national board examinations were triggers to inventing alternatives for some evaluators (West, Umland & Lucero, 1985). An additional factor has no doubt to do with the complexity of the task of designing such an integrated evaluation system.

This article contains a description of the Maastricht medical school assessment system. From its start in 1974 this school adopted a problem-based learning curriculum. In the absence of a national examination system the school itself is responsible for licensing graduates, and a proper evaluation system was an imperative requirement. In this paper the current state of affairs with regard to the Maastricht assessment system is outlined. The developments achieved so far are documented here. They are by no means *the* answer to the specific demands of assessment in problem-based learning. On the contrary, the evaluation system of the Maastricht medical school is in the middle of its development and has clearly still a long way to go, and the end goals will probably never be fully attained: the evaluation system is inherently dynamic, requiring continuous development, in which a "status quo" cannot be achieved. Nevertheless, this article is meant to describe the current state, some choices that were made, the rationale behind these choices, and some future prospects.

No explanation will be given of the rationale for problem-based learning itself and reference is made to the above mentioned literature in this regard.

The Maastricht medical school has a six year program with 150 students per class, divided into two parts. In the first four years the curriculum is divided into blocks of six weeks, each centered around a theme (e.g. blood loss, elderly people etc.). Basic, clinical, and behavioral sciences are integrated by these

themes. Students work in small groups (8 to 10) on problems and tasks offered in a block-book under the guidance of a tutor, who's role is to monitor the group process. For practical and clinical skills related to the block theme students visit the Skillslab (Bouhuijs et al., 1987) where they are able to learn and practice clinical skills in a safe laboratory setting. The program is intersected by numerous short practical periods and electives. In the second part of the curriculum students rotate through about ten clinical clerkships and electives. They spend most of their time in the clinic or family practice and discuss their experiences periodically in small groups.

Students enrolling in the medical school have graduated from academically oriented secondary schools. The Dutch secondary school system consists of various specialized schools, specifically allowing students to follow advanced education as in universities. Student selection to enter medical school is based on a national system of weighted lottery, in which secondary school achievement and chance are combined (Wijnen, 1978).

Characteristics of an evaluation system for problem-based learning[1]

The key problem in designing any evaluation system is to make the evaluation procedures congruent with educational and instructional principles. There is no stronger stimulus for a student's approach to his study than examinations (Newble & Jeager, 1983; Newble & Entwistle, 1986; Frederiksen, 1984). An examination system can be viewed as a set of rules to which students respond with strategic behavior to optimize their chance of success. Examination rules may be conceived in this perspective as "behavioral stimuli", and they may consequently be used to achieve the desired behavior. An examination system can therefore become a strong educational tool. As a consequence, whatever sophisticated instructional design is used, if the evaluation program is not in accord with the intentions of the specific instructional design, then no particular effect can be expected from this design.

In problem-based learning a number of specific requirements are to be met by an evaluation system. The first and probably most important demand concerns the principle of *self-directed learning*. In the educational system a student should learn to be fully responsible for her/his own study actions. An evaluation system prescribing for students what to do would be incompatible with the intention of this principle. Second, the *learning through practice* principle and the emphasis on *abilities beyond knowledge* poses specific demands. The integration of theory and practice, the application and use of knowledge, e.g. to solve problems, should be respected by the evaluation program. Finally, the *integration of disciplines* or subject matter is an important demand of problem-based learning.

[1]The terms assessment, evaluation and examinations will be used interchangeably, meaning all activities within the context of assessing student achievement.

To start with the last point, the medical school accepted the fact that the educational model did not permit a conventional examination system in which individual departments or faculty are responsible for their own examinations in their own disciplines. This discipline-oriented strategy would unequivocally elicit a discipline-oriented approach to learning by students. An important step was taken by *centralizing* all assessment activities. A project was established assigned to devise, maintain and investigate an evaluation system for the entire medical school adapted to the needs of problem-based learning. A number of faculty from various departments collaborate on this project on a (near) full time basis. The organizational characteristics and costs involved are dealt with in more detail later in this article.

A set of evaluation principles or goals was formulated, some of these were derivations from the educational rationale; some were general standpoints on evaluation from a broader educational perspective:

1. *The assessment should be congruent with the educational principles of the curriculum.*
   This principle is self evident and is discussed above. The rationale of the assessment system should be in concordance with the educational goals, otherwise these goals can never be reached.
2. *The assessment system should be comprehensive.*
   Assessment should not be restricted to the conventional evaluation of mere factual knowledge. The competencies which are stressed in the educational program, such as process variables or practical skills, should also be evaluated and 'rewarded'.
3. *Assessment should be a continuous process.*
   Evaluation, as is learning, is an ongoing naturally occurring activity. It should start directly at entrance in the curriculum and be repeated on a frequent basis throughout the study period. As a consequence, the rich material gathered from continuous assessments should be the basis for decision making. Reliance on momentary information (e.g. as for final examinations) seems to be an inadequate or at least incomplete use of information.
4. *Assessment should serve both summative as well as formative purposes.*
   Examinations should not be restricted to administrative decision making, but it should also serve educational needs and demands. Some authors consider examinations to be "necessary evils" (e.g. Feletti, 1980) and perhaps this is at least partially true where decision making for student promotion or *summative* assessment is concerned. However, an essential aspect of evaluation should be its "mirror" function for the student (and teacher). Tests and examinations must serve the purpose of giving feedback on the progress of a student's learning. This *formative* function is often overlooked, but should be recognized explicitly in the design process of an evaluation system or set of assessment tools. Tests in this regard serve as learning resources and there is no more "evil" in them than there is in other learning resources.
5. *The roles of teacher (e.g. tutor) and examiner should be separated in formal assessments.*
   Evaluation may be formal (structured, instrumental, quantified etc.) or informal (unstructured, verbal, casual etc.). The position was taken that the

role of a formal evaluator is incompatible with the role of a teacher (in its broadest meaning) and should be separated where possible. For example, evaluation sessions within tutorial groups are of eminent importance. However, when this is used for formal summative purposes, interpersonal relationships take on a different meaning, altering the social structure of a tutorial group. Separation of these roles can give educational involvement full rein.

6. *Instruments should satisfy all of the general demands of tests such as reliability, validity and acceptability.*
This principle may sound obvious, but in practice these demands are often not met or even considered. The mere introduction of, for instance, a test supposed to measure problem solving is clearly not enough. Before an instrument is used for formal evaluation (and certainly for summative evaluation) the characteristics in terms of reliability, validity and acceptability should be known, at least as far as possible. Added to this list of criteria should also be the effects that an instrument has on learning. As mentioned above this is a central issue for problem based learning, as it should be for any teaching method (Gronlund, 1971).

It was decided to separate aspects of clinical competence into four types of competencies which were to be evaluated distinctively: knowledge, skills, problem-solving, and attitudes. This article will describe the instruments used in the Maastricht program, formal as well as informal, to assess these competencies. The reasons for the choices that were made will be identified, some experiences will be described and some empirical findings on reliability and validity will be reported. Subsequently, a few other typical characteristics of the assessment system will be discussed: the test construction cycle, student counseling as part of the assessment program, and some organizational characteristics. Finally, the program will be evaluated against the above principles, in order to discuss (un)attained goals of the assessment system.

Besides the description of the program, the general intent of this article is to demonstrate the benefits derived from centralizing assessment activities: an assessment program may be more elaborate, can have careful quality controls, may initiate research programs regarding its instruments and procedures, etc. It enhances a rational, scientific approach, which is fully in line with the general philosophy of problem-based learning. The description below, hopefully illustrates this approach. In the assessment program some very distinctive choices were brought into practice, and whether the reader chooses to accept or reject them, their presentation here will at least give rise to a discussion of the subject, hence consistently following the rational approach.

Knowledge

In publications on assessment in the context of problem-based learning, measurement of knowledge and its associated written measurement instruments are not highly regarded. Some call these instruments a "plague" or "a pervasive thirst on the part of faculty for quantification" (West et al., 1985, p. 145)

others call them irrelevant, trivial and artificial (Pickering, 1979). These opinions, however, do not give justice to the role of knowledge. From ample research on problem-solving it is quite clear that knowledge plays a crucial role (e.g. Bransford, Sherwood & Vye, 1986; Glaser, 1984; Norman et al., 1985). The most telling critique by the adversaries is probably not the (in)significance of knowledge, but the exclusive reliance on it in many evaluation systems, combined with the educationally unwanted effects of knowledge oriented examinations. An often mentioned unwanted effect is the evocation of rote memorizing strategies of learning and the exclusive focus on details. The challenge therefore is not to discard knowledge from the evaluation program, but to overcome its unwanted side-effects. In the Maastricht assessment program two instruments are used to assess knowledge.

Block Tests

At the end of each block period all students take a test which contains questions related to that block. Questions are in the (multiple) true/false format and each test consists of 150 to 250 items. The choice of an objectively scorable type of question was motivated by the fact that knowledge is most efficiently measured using these questions. The true/false format was preferred over multiple choice, since the latter type is more difficult to construct. Although true/false questions are somewhat less reliable than multiple choice questions (which can be compensated by the fact that true/false questions are easier to construct and are more efficient with regard to testing time required), they have been proven equally valid (e.g. Norcini et al., 1983, 1985).

The content of a Block Test reflects, as closely as possible, the goals for the specific block, although that here the first problem arises. In problem-based learning students are to determine what is to be learned on the basis of the problems discussed in the tutorial group. As a result, there is a wide variation in learning paths. A test administered at the end of an academic term, made by faculty on the basis of *their* ideas and appraisal of the educational goals of that term, would forcefully direct students towards these goals. In addition, it would indeed elicit a memorizing approach to detailed (and later forgotten) material, and would change the student-centered approach to a teacher-centered one.

To overcome these side effects two measures were taken. The first one is that for each item a question mark option was made available. The student was then free *not* to answer the question, for which he will not be punished. To correct for guessing the scoring rule is the number of correct answers reduced by the number incorrect, while the question mark answers are scored neutral. A student is encouraged to use question marks whenever they were not familiar with the content of the question. It is stressed that realizing what you do not know is as important as what you do know. The second, more effective measure taken, was that students could not fail Block Tests. The results on these tests could not be used to judge negatively over progress, but conversely they could be used to compensate results on other tests of knowledge. Thus Block Tests primarily serve a formative function: they supply students with feedback on the knowledge gained from the previous learning period without forcing them to adept "hostile learning strategies".

The experience shows that the measures taken to prevent the side effects are effective. Most students do not specifically prepare for their Block Test and very few "cram" to achieve a good result. It appears that students who persist in a "cramming strategy" for Block Tests often have a generally deficient learning strategy focussed on short term memory. They have difficulty in integrating their knowledge and fail to achieve an overview of their learning material (Verwijnen, 1987). Students are generally satisfied with Block Tests. They value the feedback and feel that they can use this test for educational purposes: to see how well they did without unduly stressing themselves.

To enhance the formative value of the Block Test extra consideration is given to the way test the results are presented to students. The questions are constructed by the planning-group (a multidisciplinary group responsible for the content of the entire block period) and reviewed by several people other than the authors (the general procedure for reviewing will be described later on). For each Block Test a table of specifications (blueprint) is made by carefully translating the goals of the block into disciplines or themes involved. On the basis of this blueprint computerized profile scores are sent to all students. In addition, data on individual performance is supplied in relation to the performance of the tutorial group the student has attended and to the performance of the total class. The student thus is able to receive detailed insight in his knowledge base useful to making his own judgments for possible remedial action.

Informally block evaluation of knowledge takes also place within the block books and tutorial groups. The block books include self-evaluation sections with a number of questions, which can be used for individual or group assessments.

Progress Tests

Since summative Block Test are judged to be harmful to the educational approach, new testing ways had to be explored. To prevent a steering-effect, a solution was sought to break the direct coupling between the previous educational program and the content of a test: the so-called Progress Test was introduced. A Progress Test can best be conceived of as a kind of repeated final examination. It contains many questions (about 250 to 300, again in true/false format) together forming a knowledge sample from the entire medical cognitive domain. It represents the end objectives of the curriculum. This 'final examination' is given four times per year to *all* students in the medical school. Each test is newly constructed according to a fixed blueprint based on the International Classification of Diseases (ICD) as to yield parallel medical content. So the same test is given to all students, irrespective the class they are in, at the same time. This is repeated each three months with a test made up of new items, parallel in content to the previous one. As with Block Tests, if a student does not know the answer to an item a question mark option may be used and the same calculation rule is applied (correct minus incorrect). Naturally a freshman student will not be able to answer many questions, a second year student more and students before graduation answer the most questions correctly. What results can be seen in figure 1, in which the mean correct score per

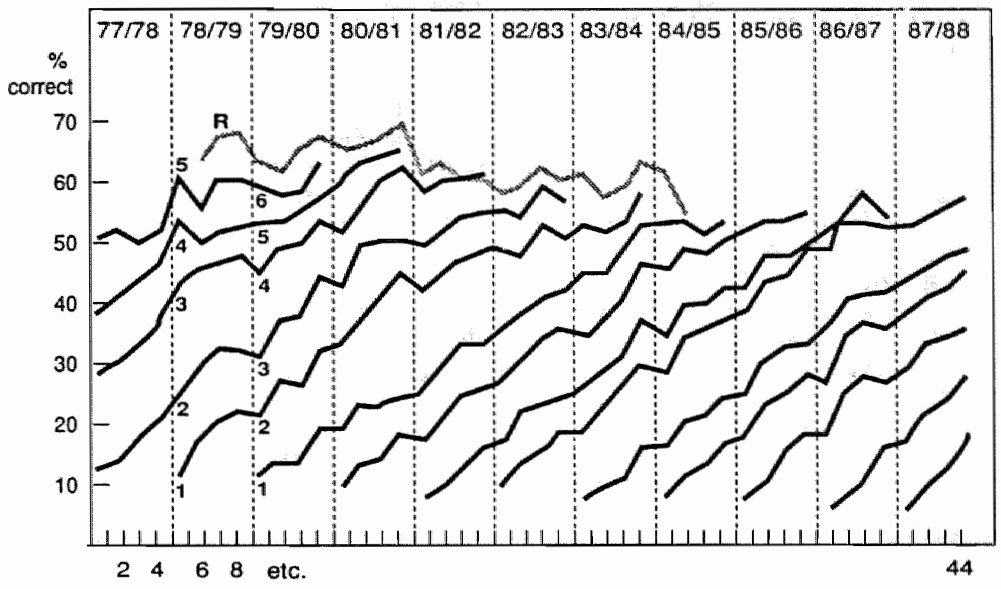

Figure 1: Mean Progress Test results across a number of cohorts, classes and years of students (continuous lines) and reference group of physicians (R; shaded line).

year-group on the Progress Test is depicted from 1978 up to and including 1988 (44 administrations in total).

In 1978 only 4 year-groups were enrolled in the program. From 1978 to 1985 a national sample of recently graduated physicians also completed each test to serve as a reference group. The top shaded line represents their average scores. The reference group was included to have some estimate of the required performance of students before final graduation, and to assess the quality of Maastricht graduates (cf. Verwijnen et al., in press). The reference group also rated each Progress Test item on relevancy as to what a graduating physician should know. This has been used for item-analysis to assess test-quality. Figure 1 shows that different year-groups achieve different scores and that each cohort consistently grows towards a final level nearly approaching the performance level of the reference group.

The Progress Test circumvents the disadvantages mentioned earlier. For an individual student it is impossible to prepare specifically for the test, since the exact content of the next test is unknown. In addition, the test covers a wide range of knowledge, so what should one prepare? On the other hand, individual learning paths are rewarded; the test does not pose any restrictions on the sequence of learning. So far experience has shown that most students indeed do not prepare themselves specifically (Verwijnen, in preparation). Preparation is in most cases restricted to global reviewing of the material studied in the previous period. A favorable effect is that this also significantly diminishes the

personal stress which is generally experienced with examinations (Harden, 1979).

By breaking the direct relationship between recent educational program and the test a key requisite of assessment within problem-based learning is realized: students are not steered by the test, whereas their individual learning paths are not interfered with. However a number of additional advantages were realized by using this approach. Some directly favorable for problem-based learning, others are general benefits; some were planned, others became apparent with growing experience. These benefits are:

- *Emphasis on recurrent relevant knowledge:*
  As contrasted with, for instance, mastery learning (e.g. Block & Anderson, 1975) problem-based learning stresses the importance of long-term memory, and integrative knowledge. Whereas learning in the mastery model is based on short well defined topics (which once learned are to be mastered), in the problem-based model learning is always directed at understanding, and integrating knowledge in the context of a problem or task. In the latter model less emphasis is placed on memorizing and rote learning, and more emphasis is placed on a "deep learning strategy" (Newble & Entwistle, 1986). Within the Progress Test attempts are made to avoid the use of items with small particular facts which can only be correctly answered when recently memorized, or which have no direct relevance for non-specialized physicians. Unlike in conventional examination programs, Progress Tests recurrently assess the same knowledge. In conventional programs, examinees rotate from examination to examination, and once having mastered and passed the examination, its content is never (or hardly ever) repeated.
- *Negative results do not necessarily have to be translated into immediately negative decisions and study-delay:*
  Progress-testing is pre-eminently a procedure of continuous evaluation. Unlike conventional examinations with re-takes, a negative result on an individual Progress Test does not have to be sanctioned immediately. The student can take this result as a warning and can try to improve his achievement the next time, instead of being forced to re-take the examination, combined with the delays involved. Through the continuous monitoring system of Progress Tests re-take examinations have thus become superfluous.
- *Progress Tests are a rich source of information:*
  The availability of comprehensive information within tests combined with frequent parallel testing over time is a rich source of information. Every three months all students receive a complete overview of where they are, related to the end objectives of the curriculum, on a broad range of topics and disciplines. After each administration they receive an overview with profile scores on all kinds of possible subdivisions within the test, such as organ system categories, disciplines involved, basic, clinical and behavioral sciences. Their individual score is related to the performance of the total class. This recurrent "X-ray" of a student's knowledge base is an important aid to identify weak spots. Directly after completion, the Progress Test booklet and the answer-key can be taken home by the students. For each test question a literature reference is supplied. All this provides for extensive information, feedback, and enhancement for further learning.

An analogue overview is made for total mean achievement on disciplines of every class which is send to the departments. They can monitor progress and evaluate growth of knowledge at certain points in time in which an emphasis of their discipline in the curriculum occurs. This can be compared with the overall proficiency level obtained from the reference group.
Finally, every item author receives feedback through the year-group scores on their item, combined with the relevancy rating from the reference group (if available).
It is also easily possible to conduct specific program evaluations with the information from Progress Tests. One can always identify subtests having to do with the target program or course to be evaluated and one can identify its complement to serve as a control condition: a subtest unrelated with the target subject. In addition, a number of pre-tests and post-tests are always available.

The construction of the Progress Test is based on a constant flow of new items written by individual faculty members from all departments. After a reviewing process and test-administration, a number of quality thresholds assessed for each item, together with a set of key-words and statistics, is stored into a computerized item-bank, ready for re-use after a number of years. Re-use of items is still limited, but with a growing item-bank (at present about 15,000 items) fewer new questions have to be made. Moreover, with each administration more information on the quality of an item is gathered, thus increasing the overall quality of the test. Together with the Block Test questions, this item-bank becomes an elaborate data base which in the future can be used for automatic computerized informal and formative assessments, whenever the individual student wishes to do so.

Progress Tests are used on a summative basis: in decision-making about the competency of knowledge, Progress Test-results thus determine marks. Cut-off scores are based on the correct-minus-incorrect scores depending on the achievements of the total group. To explain the total decision making process would be beyond the scope of this article. Its most essential feature is that decision making is based on a norm-referenced perspective of test scores[2].

The general experience on progress-testing over last the ten years is positive (Verwijnen, in preparation). Knowledge is being measured in a relevant way to a qualitatively high level and without violating the intended instructional principles. Progress Tests provides faculty members, departments and the medical school with vital information about "their performance", decision making on students can be based on multiple moments of evaluation and by the use of the item-bank high quality and reduction of cost are both achieved.

---

[2]Contrary to a domain-referenced perspective, test scores in a norm-referenced perspective are given meaning by comparison with other examinees (e.g. reference group). In a domain-referenced perspective test scores are interpreted in an absolute way: examinees should master previously defined content areas to a certain standard (defined by teachers). This rather prescriptive procedure can even be conceived of as conflicting with the educational premises of problem-based learning -- at least within the domain of knowledge --, because the model explicitly tries to circumvent learning prescriptions from the staff.

Progress testing is not unique: a similar method has been developed at the University of Missouri-Kansas City (UMKC) called the Quarterly Profile Examination (QPE) (Willoughby & Hutcheson, 1978; Willoughby, 1980). The QPE is also a comprehensive exam in a multiple choice format, also used in an innovative medical school setting (not problem-based), but only for formative purposes. Both QPE and the Maastricht Progress Test were developed independently.

Reliability and validity

Most of the research in reliability and validity on the Progress Test is summarized in Imbos (1989). Some key findings are:

The reliability of the test scores appear to be acceptable: based on the whole student body (all six years) Cronbach's alpha ranges from .98 to .96 for the correct score and from .91 to .85 for the correct minus incorrect score. The reliabilities within separate year groups are somewhat lower but still acceptable.

Validity research carried out so far on the Progress Test is positive. The tests are able to discriminate between students of different levels of ability (Imbos, 1989); they appear to be specific in that they are able to show differences between medical schools (Imbos et al., 1987); whereas differences between levels of expertise are not due to aging (Fokkema, 1986). The Progress Test has also shown to be sensitive to specific teaching effects (Imbos, 1982, 1989), and they correlate positively with other measures of medical knowledge (Stalenhoef et al., 1985). Correlations between Progress Tests over time consistently display an orderly pattern (a simplex structure; Sprooten, 1984). There is some evidence that the Progress Test is less able informative at the lowest ability levels, e.g. lowest scoring freshman (Imbos, 1989).

Skills

The emphasis within the curriculum on the learning of a broad variety of technical and clinical skills should also be reflected in the assessment program. Students spend about two to four hours a week in the Skillslab, training physical diagnostic, therapeutic, laboratory and interviewing skills. They practice on each other, on training models, simulated or real patients and are guided by special Skillslab faculty. The goal of the Skillslab is to prepare students fully on these skills before their entrance in the clinical period in years five and six. In these last two years students can still repeat, or practice in a 'safe' way, in the lab. Some clinical clerkships have a combined clinical and Skillslab program.

Skills Test

Clearly, the most valid way to assess these skills is to have students actually demonstrate their abilities. If this is combined with standard test-taking conditions and measures to enhance objective judgments, then reliable assessment should also be achieved.

After a number of trial-examinations, the so-called Skills Test was introduced in 1982. It was modeled after the Objective Structured Clinical Examination (OSCE) (Harden & Gleeson, 1979), however unlike a regular OSCE, the Skills Test is completely based on 'hands-on' performance and contains no written parts.

The Skills Test is administered in all classes (also in year 5 and 6) once a year. A single test consists of between six and twelve stations selected on the basis of a blueprint[3], together forming one circuit of two hours. An individual student rotates through the circuit, demonstrating a different skill in each station. Performance is registered by an observer, a faculty member who is generally a specialist in the content of the station involved and specifically trained for his role. The observer scores the performance on detailed checklists, on which the required task is operationally itemized. The checklists are based on standards which are used for training in the Skillslab (Lodewick & Gunn, 1982). As a check on inter-observer reliability a number of co-observers also rotate through the circuit and independently co-score the performance of a student. To test one entire class of 150 students two days are needed. Each circuit of stations is replicated four times, to allow concurrent testing of more examinees. For security reasons a new circuit is arranged each two hours, containing other stations parallel in content according to the blueprint. The content of the Skills Test is accommodated to the educational program of each class. In freshman tests the skills are still very basic, e.g. bandaging, taking blood pressure, simple interviewing and so on, but rapidly skills become much more complex. For instance, a second year student already has to conduct an extensive gynecological examination or a chest examination. Most skills are embedded in the context of patient problems, especially in higher year-groups, where they are intended to increasingly resemble real patient-doctor contacts. Although data-interpretation, differential diagnosis, and decision-making are part of the checklist, the emphasis is on the process of the skill and its technical and psychomotor procedures.

The manner in which the Skills Test is composed is analogous to the procedure of the tests described above: faculty construct stations and checklists, these are reviewed according to a standard procedure, administered in a test, the statistical information is gathered and detailed feedback is supplied to all people involved. The total bank of stations is now about 500 and re-use of stations is still limited.

The test has a summative meaning for student promotion. Station scores are calculated by calculating the percentage of correctly scored items on the checklist. Cut-off scores are primarily determined through a norm-referenced procedure, but examinees with a mean score of 70 percent correct or above automatically pass.

---

[3]The blueprint of the Skills Test is content oriented and contains categories such as neurological skills, gynecological skills, laboratory skills, interviewing skills, etc.

After six years of use, the Skills Test is now well established within the evaluation program. Faculty and students agree that the test is a valid way to assess the skills intended and they value the feedback (Van Luyk et al., 1986).

An apparent disadvantage is the organizational complexity and the logistics involved with skills-testing. Naturally, performance-based instruments are more demanding compared to paper and pencil tests, but on the other hand this should not be overrated. Williams et al. (1987), Stillman et al. (1986), calculated that costs for performance-based assessment methods are well within reasonable limits. With growing experience and some automation it is our experience that logistic problems can be controlled. Clearly, the centralized approach of the Maastricht evaluation system is beneficial in this regard.

In the first four years, knowledge pertaining to skills is assessed by the Block Test. As mentioned earlier, the Skillslab program is interwoven with the educational block program. For formative purposes, part of each Block Test is reserved for questions concerning skills.

Informal evaluation of skills is an inherent part of the training of skills. Training is mainly carried out in small groups under the supervision of a Skillslab teacher who structures evaluation sessions. Simulated patients are often used and they are trained to supply feedback, not only on the social dimension, but also with regard to physical examination. For instance, a number of women volunteer as patients for gynaecological examinations. Their evaluation of a student's actions are crucial for the proficiency of students in these skills.

Reliability and validity

Generalizability analysis of Skills Test-scores has shown that the major source of measurement error comes from the low correlation of examinee performance from one station to the other (Van der Vleuten, Van Luyk & Swanson, 1988). This measurement error is a much larger source than the error due to differences between observers. Observer agreement is generally very acceptable and rather comparable across content areas. Lowest inter-rater reliability is found for interviewing skills (Van der Vleuten & Van Luyk, 1987). Training of (physician) observers seems not to influence the accuracy of ratings very much, and being an expert in the tested field as a requirement is not as imperative for accurate scoring as one might intuitively suggest (Van der Vleuten et al., in press).

The limited consistency of performance across stations implies that relatively many stations are needed to yield reliable scores. Minimal reproducibility is reached with approximately 15 stations, or an overall testing time between three and four hours. To improve the reliability of the Skills Test alternative testing strategies (e.g. sequential testing; Swanson & Norcini, in press) are under study.

Results on the validity of the Skills Test in the studies carried out so far are positive. Scores on the Skills Test converged to general judgments of observers (Van der Vleuten & Van Luyk, 1986), to a written knowledge test of skills (Van der Vleuten, Van Luyk & Beckers, 1989) and appeared predictive for later clerkship performance (Van der Vleuten, Van Luyk & l'Espoir, 1987).

Problem-solving

Problem-solving is a core concept in problem-based learning. In the educational model students are taught (or taught by themselves) to become self-learning competent problem-solvers. Consequently, problem-solving should be equally represented in the evaluation system. However, this is not currently the case with the Maastricht evaluation system. Although numerous informal assessments take place, formal assessment of problem-solving is restricted to the last two years of the curriculum, and on a rather weak basis through the use of clerkship ratings for assessment of clinical reasoning. Part of the explanation for this absence was mentioned in the introduction, relating to the stage of development of the evaluation system. However, it would have been relatively easy to introduce formats developed elsewhere. The supply of techniques on the market is rich: among others, Patient Management Problems (Rimoldi, 1961; McGuire & Solomon, 1976), the Triple Jump Exercise (Powles et al., 1981), Modified Essay Questions (Feletti & Engel, 1980), Portable Patient Problem Pack (Barrows & Tamblyn, 1977), Case-reports (McLeod, 1987), Clinical Reasoning Test (Williams et al., 1987), computer simulations (Norcini et al., 1986; Taylor et al., 1976). Common denominator in these instruments is that they confront examinees with (written) clinical simulations. However, critical appraisal of their empirical achievements in studies of reliability and validity yields clear warnings. In their review of reliability and validity studies, Swanson et al. (1987) conclude that there is little evidence these measures provide unique measurement information and they suggest:

> *'While clinical simulations pose some interesting research challenges, until research efforts improve their psychometric characteristics, they should probably not be used.'* (p. 23).

In part, these disappointing results are probably due to the fact that the quintessence of the problem-solving concept is still rather unclear, and that more fundamental (psychological) insight is needed. This notion has also affected the research in the medical problem-solving area in the last decade. A shift can be noticed from psychometric studies of instrument applications to more fundamental cognitive psychological research, such as the study of expert-novice differences (e.g. Boshuizen, 1989). However, much more work in this young scientific discipline has to be undertaken before the results can be applied to practical construction of concrete measurement instruments, although some first attempts to bridge the gap between cognitive psychological research and measurement of problem solving have recently been made (Norman, 1987). To conclude, a major reason for the absence of instruments for the assessment problem-solving assessment in the (formal) Maastricht assessment system lies in the under developed state of the art of present day science in the area.

This should not be taken as an alibi or excuse to lean back and wait for what may come. On the contrary, an important task for every professionally involved medical educationalist in assessment is to follow actively and participate in the attempts to improve measurement of problem-solving. Research information unequivocally points to the seriousness of the problem of "content specificity" of problem-solving: achievement on one problem is not very pre-

dictive for solving another problem within the same context (e.g. Norman et al., 1984), thus requiring tests that contain relatively many problems or cases and yielding to (unacceptably) long testing time requirements. This has led researchers to seek more efficient testing methods such as focussing test material on the central characteristics of a problem or its key feature (Norman et al., 1985; Bordage et al., 1987). Active research in this area is also carried out in Maastricht (De Graaff et al., 1987).

Instruments for problem-solving have not (yet) passed the demands of reliability and validity, but they may still be useful for educational purposes. For the purpose of formal assessments they should be handled with caution, but for informal use one might profit from the educational effect of these instruments. Some of the instruments mentioned above are voluntarily used by students in their tutorial groups. Recently, first attempts are made to introduce computer programs and simulations, so as to exploit the new possibilities of the computer.

Additional important informal assessment moments in this area are the simulated patient contacts. In the first four years of the curriculum each student has a simulated patient contact in the Skillslab about each three weeks. This encounter is videotaped and seen by two faculty members: a physician and a behavioral scientist. This encounter is discussed in group sessions. Sometimes a written case report from the student is requested.

Clinical Ratings

Standardized clinical ratings in the clinical rotations are used as formal assessments of problem solving or clinical reasoning[4]. These ratings are completed at the end of each clinical period by the clinical supervisor of the student. The ratings are based on a number of sub-divisions which are considered important for problem solving (data-gathering, data-interpretation and data-management). Ratings are recorded on a three point scale and ample space is supplied to give the supervisor the possibility to enhance his rating with verbal comments.

This conventional assessment strategy has its conventional drawbacks as documented in the literature (O'Donahue & Wergin, 1978; Levine & McGuire, 1971; Tonesk, 1983): the personal influence of the rater, the global nature of the rating, questionable validity, etc. The Clinical Rating is also the one exception in the assessment system in which the principle of separation between "assessor and teacher" is disregarded.

Attitudes

Generally the same situation as with of problem-solving applies to attitudes. Although attitudes are considered important, no formal assessments exist for the same reasons as mentioned above, with one exception. The Clinical Ratings in year five and six contain a rating on attitudes. However, the term is not specified and the rating is very global as a result.

---

[4]They also enclose ratings of knowledge, attitudes and general impression ratings. However they are the main source of information for clinical reasoning.

Informal explicit assessment of attitudes mainly takes place in two situations. The first one is the above described simulated patient contacts in the Skillslab. Students receive information on how they relate to the patient, are stimulated to be aware of their functioning with regard to this specific (medical) problem, etc. This often leads to more elaborate discussions between students on these topics (in the context of their interviewing skills training; cf. Van Dalen, Zuidweg & Collet, 1989). Secondly, in all year groups attitude discussion groups are organized. Students have special group meetings and discuss (medical) experiences and their significance to their personal development towards becoming a physician.

Thus far a number of instruments pertaining to the diverse competencies have been outlined. For each of these, the underlying rationale and description was delineated. Clearly, the system is far from complete and with no doubt for some too cautious. In the next paragraphs other key characteristics of the evaluation system will be discussed. These characteristics surmount individual instruments, but in our view are essential for the functioning of the total evaluation program.

The test construction cycle and quality control

Experience has shown that the quality of test material is improved significantly by a number of fairly simple measures within the construction process of a test. Therefore for each of the instruments in the assessment system a standard construction procedure has been introduced. Key elements in this procedure are: item-reviewing, student comments, test-statistics and computerized storage and retrieval. This procedure is the responsibility of a special test-committee and for each instrument such a committee has been established. The production cycle of a Progress Test will be described here as an example, but the procedure for Block Tests and Skills Tests are essentially the same. However, the most experience exists with regard to the Progress Test and preliminary data are available (Hessen & Verwijnen, 1987).

The cycle begins with a constant flow of test-items made by faculty of all departments. These are gathered in a so-called item-pool. Before test-administration, items are drawn from the pool, stratified according to the categories of the test blueprint, but random over disciplines (departments). As a result the chance of any item being drawn for a certain discipline depends on the stock in the pool. The departments regularly receive frequent surveys of their stock to serve as stimuli for the construction of new items. In addition, each department is allotted specified "education time" of half an hour for each question which is

finally included in the test.[5] The selected items from the pool are than reviewed by the Progress Test review committee.

This eight member committee is composed of faculty members from across the disciplines. They check the content, the wording and they judge the relevance of each item. If the committee considers alterations necessary, communication between author and committee is followed. After the reviewing process, items are administered in the test. The students are encouraged to supply comments on items within a few days after administration. This material, together with item statistics is again reviewed by the committee. Communication with authors is repeated for disputable items. After elimination of some of these (generally 5 to 10 percent of the whole test), definitive results are determined. All items are then entered into a computerized item-bank for later re-use.

The essence of this process is very simple: evaluation material is scrutinized by several people (including students) other than the author. The precise effect on the quality of the resulting evaluation is hard to measure. However, the statistics on the number of alterations are rather imposing (Hessen & Verwijnen, 1987): in the entire process 75 percent of all questions are altered in one way or another. From this 75 percent alteration was either due to ambiguous wording (60 percent), or to disputable content (25 percent) and to questionable relevance (12 percent). The fact that three quarters of original drafts have some kind of detectable flaw when thoroughly inspected is remarkable and implies that the reviewing process probably is an important procedure for improving the overall quality of testing. It also gives us clear warnings about unreviewed tests, as is the case for many examinations in conventional educational systems.

The construction cycle reinforces itself. The evaluation material finally stored has passed a number of quality thresholds. The re-use of this stored material then is not only an economic strategy, re-administration of the evaluation material will again cumulate information, and improve the test quality. Moreover, the statistical information can be applied to optimize the psychometric characteristics of the test. For instance, the critical intercase reliability of the Skills Test may be improved by not re-using stations (cases) having low intercorrelations with other stations. The re-use of testing material however is only possible if the itembank is large enough to prevent examinees from memorizing old material.

Student counseling and decision making

An assessment system should be more than the sum of its constituent instruments and their resulting scores. Whatever reliable and valid measures one implements, whatever conscientious procedures one installs, evaluation requires

[5]The medical school has a token system for all educational activities. For each educational task a fixed number of hours is determined. For instance, 40 hours is the credit for being a tutor, being an observer in the skills-test 10 hours, etc. The summation of this labeled time per department should match the overall department staff labeled on education.

more. It should not be restricted to routinely computational procedures and actuarial decision making. The literature on predictive characteristics of testing instruments gives articulate warnings (e.g. Wingard & Williamson, 1973). An effective evaluation system requires an additional (human) contribution to augment the bare flow of assessment information to each individual student.

With this purpose in mind a student guidance system has been introduced very explicitly as an integral part of the examination program, in which students enrolling in the medical school are allocated to a faculty member (who is therefore compensated by the token system). To become a student counselor a staff member must have had a number of years of experience in the medical school and he must have attended a counseling-course in the faculty development program. Counselors meet each student a few times per year. The personal student file is the basis of these sessions, and contains all test results of a particular student and all other information a student wishes to store (personal records, case-reports etc.). Both student and counselor analyze the current state of affairs: 'Was there enough progress; were problem areas encountered; was the student's study strategy effective; are there any study-topics which need more attention' are examples of questions which might be discussed in these counseling sessions.

The formal decision making over students is the responsibility of the Certification Committee of the medical school. Most of these decisions are based on straightforward applications of a number of rules pertaining to the test scores. However, explicitly included in the examination rules is that the counselor and student compose an advice, or better "contract", based on the test material and other information resulting from their session(s). The contract is signed by both and sent to the Certification Committee. It constitutes the basis for final decision making. The advice has special significance in cases where test scores indicate borderline performance. Specifically included in the combination rules for test scores, is that the end result of this combination may be indecisive. It actually means that the test results have not supplied *enough* information to take a final decision. The information from the counseling sessions, and the counselor's advice may supply the additional information needed to take that final decision.

It should be stressed that the emphasis of the counseling system is not exclusively on decision making, but on helping the student through the study. The detailed feedback from the test results has appeared to be helpful for this purpose. Although further research has to be done, a number of case studies have demonstrated that the profile scores on tests, in combination with the student's specific history, results in relevant and valid information, making early detection of students at risk possible (Verwijnen, 1987).

Organizational characteristics

As delineated above, the medical school has chosen a centralized approach to the development of an assessment system by installing a group of faculty responsible for this development. This choice was motivated by earlier ex-

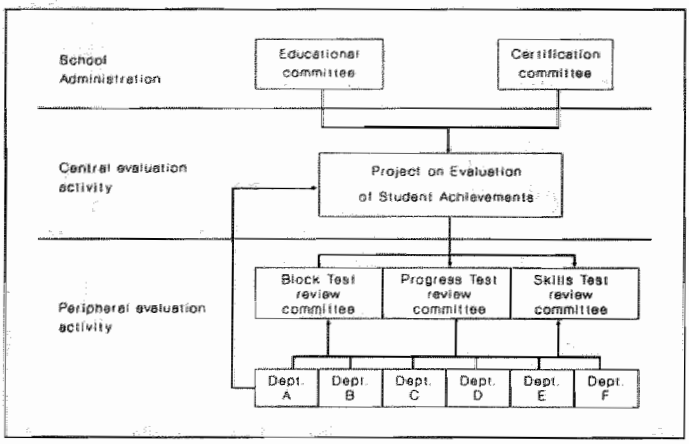

Figure 2: Organizational scheme of evaluation activities.

periences with a less centralized system. Originally, several groups were accountable for assessment. According to the matrix organization system of the school (Bouhuijs, in press) several project groups were originally assigned separate evaluation tasks, e.g. separate assessments for formative and summative evaluation. This approach led to problems of coordination, sometimes to poor quality of test material, and to a problem in the continuity of the tasks involved, caused by the frequent staff rotations. In 1982 the educational committee responsible for the educational program and the certification committee responsible for certification, decided to integrate all efforts into one project with faculty whose main task became assessment. The resulting organizational scheme is depicted in figure 2.

The total scientific staff on the project is approximately 4 full-time equivalents (FTE) of which 2 FTE is labeled for research. The project has about 6 FTE administrative-operational staff at their disposal. The costs are accounted for by the compensatory effect of centralization: less need exists for "peripheral" staff in assessment. Staff from the central project are coordinators of the review committees, but other committee-members come from regular departments. Membership on a review committee is a four year term, compensating regular educational time by the token system. As mentioned before, test material is produced by individual departments. They are, within the procedures described, responsible for the content of assessment. The total peripheral evaluation activities are estimated to take about 5 FTE, making the total investment on scientific staff about 9 FTE.

On first sight this may appear as a large investment. However, it should be realized that this constitutes approximately 4 percent of the total scientific staff of the entire medical school and 12 percent of the staff needed for all educational activities. Although the assessment time involved in other (conventional)

systems is unknown, these percentages can be considered to be perhaps even relatively low.

Discussion

The importance of assessment systems in relation to educational innovations is being increasingly recognized (e.g. Newble & Entwistle, 1986). Despite this attention, descriptions of integrative assessment systems are rare. What can be found is ample documentation on specific evaluation instruments. However, a system of assessment is more than a collection of testing instruments. The instruments are very important because they are the elements or 'bricks' of the total structure, but it is the structure as a whole which is equally important. Besides some general recommendations, hardly any educational methodology exists explicitly in the planning of assessment systems.

An attempt in this paper to account for the effort in Maastricht to construct an integrative evaluation system for student achievements is made. A cardinal characteristic of this system is the centralized approach. It allows for a more rational strategy for assessment, and it is argued that this strategy is beneficial to the quality of the assessment program, and hence for the general quality of education. Integrative large scale assessment procedures with carefully built-in quality controls as described above are scarcely possible without this organizational strategy. Moreover, using this strategy, experience is more easily accumulated and specific research projects investigating the assessment program become more feasible. The centralized approach enhances a scientific approach. Not only through these specific research projects, but also because the evaluation program becomes visible and public to a(n) (scientific) audience and hence *open for debate*[6], leading to alterations and improvements.

The assessment system is still in the middle of its development. Some of the evaluation principles outlined in the introduction have been realized, others are not yet or only partly attained.

Partly attained is the *principle of congruence* between assessment program and educational premises. With the technique of Progress Testing the attempt not to steer student learning by the testing program has succeeded reasonably, and the principle of self-directed learning is respected. Unfortunately, not every instrument can be transformed into a Progress Test format. Because of its limited sampling possibilities (due largely to time limitations), the Skills Test cannot have Progress Testing-characteristics, and, as a consequence, a tendency can be observed that students prepare themselves specifically for the (expected) content of these tests by learning checklists by heart: the test starts directing learning. Through careful analysis of the content of the checklists, and through

[6]This argument is not restricted to a scientific level of debate (e.g. with colleagues in the field), but also holds within the medical school. Because the assessment program is centralized and fairly visible, it is constantly debated and criticized within the medical school at all levels (e.g. faculty, students, administrative committees, etc.). Consequently, arguments and empirical evidence (many of the specific research projects originate from these discussions) are continuously exchanged, and refinements to its rules and procedures are made.

investigating effects of more globalized and focused checklists, this effect hopefully may be reduced in the future.

The assessment system has almost actualized the principle of *integration of disciplines*. Except for the Clinical Ratings, all instruments are stratified on the basis of subject matter, and not on the basis of disciplines. Apart from this preventing (endless) discussions about the weighing of disciplines within tests (and the curriculum), it has reinforced the integrative premise of problem-based learning. Also the strategy of multi-disciplinarity has been applied successfully in the organizational structure of the people responsible for the assessment system.

The *comprehensiveness* of the assessment system is also partly attained. Although the program still relies heavily on the evaluation of knowledge, principle and empirical arguments have been put forward for this emphasis. Regardless of the instructional method involved, accumulation of knowledge is a fundamental issue in any educational system. As was mentioned before, opponents to (objective) knowledge-tests probably do not deny the significance of knowledge, but they oppose the exclusive reliance of most assessment programs on these tests and the effects they have on the students' learning[7]. The latter argument needs no further discussion, whereas the first objection does not hold for the Maastricht program. The elaborate representative assessment of technical and clinical skills has no counterpart in any other educational institution, although the relevance of these skills is widely recognized. The evaluation of problem-solving, especially in relation to the clinical rotations, is an area which clearly deserves more attention. Although numerous informal evaluation activities occur, formal assessment of this vital competency is still lacking. The same conclusion can be drawn in the area of attitudes, but from the literature it is evident that short term results cannot be expected; further research is the only action possible for the moment. These shortcomings are hopefully ameliorated by the counseling system, the primary purpose of which is to assess and adjust all evaluation material in relation to the individual case of each student.

The present evaluation system fulfills to a large extent the requirement of a *continuous* program. Assessment directly starts from the beginning, is repeated on a frequent basis, and carries on until final graduation. Even when students are in their clinical clerkships, assessment with Progress Tests and Skills Tests is part of their program. The Progress Testing format by nature is pre-eminently an example of a continuous assessment method. Probably the best demonstration of continuity is that separate large scale final examinations are not necessary, and in fact are non-existent in the Maastricht assessment program.

The *formative value* of instruments used has been outlined earlier. Where possible, profile scores are provided, verbal feedback is given, literature references are supplied, etc. A very important formative aspect of the assessment

[7]As far as the disapproval of the use of multiple choice items is concerned because of their restricted ability to test simple recall of information (and not higher-order cognitive abilities), this critique is not justified. It is not the question format which dictates the cognitive level being tested, but the task the examinee has to perform. McGuire (1987) has labeled this critique as a damaging myth (p. 49).

program is the counseling system. The purpose is to monitor the student throughout his entire study, to guide and if necessary to intervene in his educational career. The basis for all this, is in the assessment results. They explicitly have a diagnostic function in this regard. The test results are thus given individual meaning and have highly formative significance.

The principle of *separation of educational and judgmental roles* is realized for most of the assessment program. There are two exceptions. The first one concerns the clinical ratings. Here a clinical supervisor is responsible for each student's educational experiences, as well as for the rating of achievement. It was mentioned earlier, that the ratings have drawbacks, and clearly this instrument, and the clinical assessment program in general, need more attention in the future. The second exception is the student-counselor. The student-counselor explicitly has a vote in the promotion of the student, but in a strict sense this is incompatible with his formative function. Students have complained about this summative role of counselors, but their corrective human contribution to the overall promotion system has until now prevailed over the associated disadvantage of this role.

The demands of *reliability, validity and acceptability* are naturally not fully met by the assessment instruments. A few areas of attention are: the validity of Progress Tests in the lower ability regions (e.g. first year), the reliability of the Skills Test, the value of the clinical ratings and the validity of some of the general procedures used (e.g. effect of reviewing, use of test results for remedial purposes, etc.). In principle, however, this kind of research *is* possible, and will be carried out. Very likely these requirements will never be entirely satisfied, but at least it will be known where they are *not* achieved. The rationale of the overall system and the expertise of those involved may hopefully correct for these shortcomings.

References

Barrows, H.S. (1985) *How to Design a Problem-Based Curriculum for the Preclinical Years.* New York: Springer.

Barrows, H.S. & Tamblyn, R.M. (1977) The portable patient problem pack (P4). A problem-based learning unit. *Journal of Medical Education, 52*, 1002-1004.

Barrows, H.S. & Tamblyn, R.M. (1980) *Problem-based learning.* New York: Springer.

Block, J.H. & Anderson, L.W. (1975) *Mastery learning in classroom instruction.* New York: Macmillan.

Bordage, G. & Page, G. (1987) An alternative approach to PMPs: The "key features" concept. In: I.R. Hart & R.M. Harden, (Eds.), *Further Developments in Assessing Clinical Competence.* Montreal: Heal-Publications.

Boshuizen, H.P.A. *De ontwikkeling van medische expertise: Een cognitief psychologische benadering.* (The development of medical expertise: A cognitive psychological approach.) Meppel: Krips Repro.

Bouhuijs, P.A.J. The maintenance of educational innovations in medical schools. (In press) In: Khattab, T., Schmidt, H., Nooman, Z. & Ezzat, E. (Eds.), *Innovation in Medical Education: An Evaluation of Its Present Status.* New York: Springer.

Bouhuijs, P.A.J., Vleuten, C.P.M. van der & Luyk, S.J. van, (1987), The OSCE as a part of a systematic skills training approach. *Medical Teacher, 9*, 183-191.

Bransford, J., Sherwood, R. & Vye, N. (1986) Teaching thinking and problem-solving. *American Psychologist, 41*, 1078-1089.

Dalen, J. van, Zuidweg, J. & Collet, J. (In press) The curriculum of communication skills teaching at Maastricht Medical School. *Medical Education.*

De Graaff, E., Post, G.J. & Drop, M.J. (1987) Validation of a new measure of clinical problem-solving. *Medical Education, 21*, 213-218.

Feletti, G.I. (1980) Evaluation of a comprehensive programme for the assessment of medical students. *Higher Education, 9*, 169-178.

Feletti, G.I. & Engel, C.E. (1980) The modified essay question for testing problem-solving skills. *Medical Journal of Australia, 1*, 79-80.

Fokkeman, F. (1986) Het gebruik van voortgangstoetsen bij het vergelijken van medische curricula: Een begripsvalidatie. (The use of progress tests in comparisons of medical schools: A construct-validation). *Heijmans Bulletin*, HB-86-812-SW. Groningen: University of Groningen.

Frederiksen, N. (1984) The real test bias: influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Glaser, R. (1984) Education and thinking: The role of knowledge. *American Psychologist, 39*, 93-103.

Gronlund, N.E. (1971) *Measurement and Evaluation in Teaching.* New York: MacMillan, 1971.

Harden, R.M. (1979) How to assess students: An overview. *Medical Teacher, 1*, 65-70.

Harden, R.M. & Gleeson, F.A. (1979) ASME Medical Education Booklet No. 8. Assessment of clinical competence using an objective structured clinical examination (OSCE), *Medical Education.*

Hessen, P. & Verwijnen, G.M. (1987) Necessity of a test review committee in test construction. *Paper presented at the International Symposium on Evaluation in Medical Education.* Beer-Sheva, Israel.

Imbos, Tj. (1982) Effecten van computersimulaties in blok 1.5. (Effects of computer simulation). *Internal Report Project on Evaluation of Student Achievements (nr. 22)*, University of Limburg.

Imbos, Tj. (1989) *Het gebruik van einddoeltoetsen bij aanvang van de studie.* (The use of end objective related achievement tests in freshman years). Maastricht, University of Limburg, Doctoral Dissertation.

Imbos, Tj., Hessen, P.A.W. van, Muytjens, A., Snellen, H., Verwijnen, G.M., & Wijnen, W.H.F.W. (1987) Some examples of the possibilities of program independent achievement testing with progress tests. *Abstracts of the International Symposium on Evaluation in Medical Education*, p.33, Beer Sheva, Israel.

Kantrowitz, M., Kaufman, A., Mennin, S., Fulop, T & Guilbert, J. (Eds.) (1987) *Innovative Tracks at Established Institutions for the Education of Health Personnel.* Geneva: World Health Organization Publication no. 101.

Kaufman, A. (Ed.) (1985) *Implementing Problem-Based Medical Education.* New York: Springer.

Levine, H.G. & McGuire, C.H. (1971) Rating habitual performance in graduate medical education. *Journal of Medical Education, 46,* 306.

Lodewick, L. & Gunn, A.D.G. (1982) *The Physical Examination. An Atlas for General Practice.* Lancaster: M.T.P. Press.

Luyk, S.J. van & Vleuten, C.P.M. van der (1987) Vijf jaren toetsing van vaardigheden. *Syllabus Symposium Onderwijsvernieuwing: Een greep uit ervaringen met probleemgestuurd onderwijs,* Maastricht. (Five years of testing skills. *Proceedings Symposium on Educational Innovations: A Corollary of Experiences with Problem-based Learning,* Maastricht.)

McGuire, C. (1987) Written methods for assessing clinical competence. In: I.R. Hart & R.M. Harden, (Eds.), *Further Developments in Assessing Clinical Competence.* Montreal: Heal-Publications.

McGuire, C.H. & Solomon, C.M. (1976) *Construction and Use of Written Simulations.* Chicago: The Psychological Corporation.

McLeod, P.J. (1987) Faculty assessments of case reports of medical students. *Journal of Medical Education, 62,* 673-677.

Neufeld, V.R. & Norman, G.R. (1985) *Assessing Clinical Competence.* New York: Springer.

Newble, D.I. & Jaeger, K. (1983) The effect of assessment and examinations on the learning of medical students. *Medical Education, 17,* 165-171.

Newble, D.I. & Entwistle N.J. (1986) Learning styles and aproaches: implications for medical education. *Medical Education, 20,* 162-165.

Norcini, J.J., Swanson, D.B. & Webster, G.D. (1983) Reliability, validity and efficiency of various item formats in assessment of physician competence. *Proceedings of the Twenty-Second Annual Conference on Research in Medical Education,* Washington: American Association of Medical Colleges.

Norcini, J.J., Swanson, D.B., Grosso, L.J. & Webster, G.D. (1985) Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education, 19,* 238-247.

Norcini, J.J., Meskauskas, J.A., Langdon, L.O. & Webster, G.D. (1986) An evaluation of a computer simulation in the assessment of physician competence. *Evaluation in the Health Professions, 9,* 286-304.

Norman, G.R (1987) Measurement characteristics of cognitive probes. *Proceedings of the Twenty-Sixth Annual Conference on Research in Medical Education,* Washington: American Association of Medical Colleges.

Norman, G., Bordage, G., Curry, L., Dauphinee, D., Jolly, B., Newble, D., Rothman, A., Stalenhoef, B., Stillman, P., Swanson, D. & Tonesk, X. (1985) A review of recent innovations in assessment. In: Wakeford, R., Bashook, P., Jolly, B. (Eds.) *Directions in Clinical Assessment.* Cambridge: Office of the Regius Professor Of Physic Cambridge University School of Clinical Medicine.

Norman, G.R. & Tugwell, P., Feightner,J.W., Muzzin, L.J., & Jacoby, L.L. (1985) Knowledge and clinical problem solving. *Medical Education, 19*, 344-356.

O'Donahue, W.J. & Wergin, J.F. (1978) Evaluation of medical students during a clinical clerkship in internal medicine. *Journal of Medical Education, 53*, 55-58.

Pickering, G. (1979) Against multiple choice questions. *Medical Teacher, 1*, 84-86.

Powles, A.C.P., Wintrup, N., Neufeld, V.R., Wakefield, J.H., Coates, G. & Burrows, J. (1981) The triple jump exercise: Further studies of an evaluative technique. *Proceedings of the 20th Annual Conference on Research in Medical Education*, Washington: American Association of Medical Colleges.

Rimoldi, H.J.A. (1961) The test of diagnostic skills. *Journal of Medical Education, 30*, 72-79.

Schmidt, H.G. (1983) Problem based learning: Rationale and description. *Medical Education, 17*, 11-16.

Schmidt, H.G. & De Volder, M.L. (1984) *Tutorials in Problem-Based Learning.* Assen: Van Gorcum.

Schmidt, H.G., Dale Dauphine, W.D. & Patel, V.L. (1987) Comparing the effects of problem-based learning and conventional curricula in an international sample. *Journal of Medical Education, 62*, 305-315.

Sprooten, J. (1984) Lisrel simplex analyses. *Internal Report Project on Evaluation of Student Achievements (nr. 22)*, University of Limburg.

Stalenhoef, B.S., Snellen, H.A.M. & Imbos, Tj. (1985) Jaarverslag bloktoesten, 1983/84 en 1984/85. (Annual report on Block Tests.) *Internal Report Project on Evaluation of Student Achievements (nr. 109)*, University of Limburg.

Stillman, P.L., Swanson, D.B., Smee, S., Stillman, A.E., Ebert, T.H., Emmel, V.S., Caslowitz, J., Greene, H.L., Hamolsky, M., Hatem, C., Levenson, D.J., Levin, R., Levinson, G., Ley, B., Morgan, J., Parrino, T., Robinson, S. & Willms, J. (1986) Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine, 105*, 762-771.

Swanson, D.B. (1988) A measurement framework for performance based tests. In: I.R. Hart & R.M. Harden, (Eds.), *Further Developments in Assessing Clinical Competence.* Montreal: Heal-Publications.

Swanson, D., Norcini, J., and Grosso, L. (1987) Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*, 220-246.

Taylor, W.C., Grace, M., Taylor, T.R., Fincham, S,M. & Skakun, E.N. (1976) The use of computerized patient management problems in a certifying examination. *Medical Education, 10*, 179-182.

Tonesk, X. (1983) *The Evaluation of clerks: Perceptions of clinical faculty.* Washington: Association of American Medical Colleges.

Verwijnen, G.M. (1987) Betekenis van studieresultaten bij studiebegeleiding. (Meaning of student achievement in student counseling.) In: De Grave, W.S. & Nuy, H.J.P. (Eds.) *Leren studeren in het hoger onderwijs* (Learning to study in higher education). Almere, The Netherlands: Versluys.

Verwijnen, G.M. (In preparation). *Student Opinions on Progress Testing*.

Verwijnen, G.M., Vleuten, C.P.M. van der & Imbos, Tj. (In press) Comparing an innovative medical school with traditional schools: An output analysis in the cognitive domain. In: Khattab, T., Schmidt, H., Nooman, Z. & Ezzat, E. (Eds.) *Innovation in Medical Education: An Evaluation of Its Present Status*. Springer Puplishing Company.

Vleuten, C.P.M. van der & Luyk, S.J. van, (1986), A validity study of a test for clinical and technical medical skills. In: I.R. Hart, Harden, R.M. & Walton, H.J. (Eds.), *Newer Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Vleuten, C.P.M. van der, Luyk, S.J. & Peet, D. (1986) The assessment of clinical and technical skills at the medical school of Maastricht. In: I.R. Hart, Harden, R.M. & Walton, H.J. (Eds.), *Newer Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Vleuten, C.P.M. van der, Luyk, S.J. van & l'Espoir, N.E.J.C. (1987), Effecten van vaardigheidsonderwijs (Effects of teaching skills). In: F.J.R.C. Dochy & S.J. van Luyk (Eds.). *Handboek Vaardigheidsonderwijs* (Handbook of Skills Education). Lisse: Swets & Zeitlinger.

Vleuten, C.P.M. van der, Luyk, S.J. van & Swanson, D.B. (1988), Reliability (generalizability) of the Maastricht Skills Test. *Proceedings of the Twenty-seventh Annual Conference on Research in Medical Education (RIME)*, Chicago, USA.

Vleuten, C.P.M. van der, Luyk, S.J. van & Beckers, A.J.M. (1989) A written test as an alternative to performance testing. *Medical Education, 23*, 97-107.

West, D.A., Umland, B.E. & Lucero, S.M. (1985) Evaluating student Performance. In: Kaufman, A. (Ed.) *Implementing Problem-Based Medical Education*. New York: Springer.

Wiliams, R.G., Vu, N.V., Barrows, H.G. & Verhulst, S. (1984) Profile of the Clinical Reasoning Test (CRT): An objective measure of problem solving skills and proficiency in using medical knowledge. In: Schmidt, H.G. & De Volder, M.L. (Eds.) *Tutorials in Problem-Based Learning*. Assen: Van Gorcum.

Williams, R.G., Barrows, H.S., Vu, N.V., Verhulst, S.J., Colliver, J.A., Marcy, M. & Steward, D., (1987), Direct, Standardized Assessment of Clinical Competence. *Medical Education, 21*, 482-489.

Willoughby, T.L. (1980) Quarterly Profile Examination, *RIME exhibition handout*, Annual Meeting of the Association of American Medical Colleges, Washington.

Willoughby, T.L. & Hutcheson, S.J. (1978) Edumetric validity of the Quarterly Profile Examination. *Educational Psychology Measurement, 38*, 1057-1061.

Wingard, J.R. & Williamson, J.W. (1973) Grades as predictors of physician's career performance: an evaluative literature review. *Journal of Medical Education, 48*, 311-322.

Wijnen, W.H.F.W. (1978) Netherlands. In: Burn, B. (Ed.) *Admission to Medical Education in Ten Countries*. New York: International Council for Educational Development.

HOOFDSTUK 2

BETROUWBAARHEID VAN OBSERVATIETOETSEN VOOR PRAKTISCHE VAARDIGHEDEN IN HET MEDISCH ONDERWIJS

Abstract

Tests for observations of practical skills are increasingly popular in many educational settings. Despite this popularity their psychometric characteristics are fairly unknown. This study reports on the reliability of an observation test which is used at the medical school of the University of Limburg. Every year all medical students show their ability in practical medical skills by means of a so-called skills-test, consisting of a series of stations in which students have to demonstrate their skills apprehended.

Medical students from all classes in the 1984/1985, 1985/1986, and 1986/1987 academic years were included in the study. Using methods derived from generalizability theory, analyses investigated interrater reliability and reproducibility of scores.

Overall interrater reliability appeared to be sufficient across all classes. On the other hand, variation in the quality of examinee performance from station to station proved to be a large source of measurement error. Decision studies indicated that, depending on the interpretation perspective taken, a minimum of three to five hours of testing time is required to obtain reasonably reproducible scores, regardless of year of training, and despite differences in test content. Some strategies for improvement of the reliability of observational tests are discussed.

Inleiding

Praktische vaardigheden spelen in vele sectoren van het onderwijs een belangrijke rol, met name en vooral in de meer beroepsgeoriënteerde disciplines. Hoewel er in onderwijskundige zin een groeiende aandacht bestaat voor praktische vaardigheden (b.v. De Klerk, 1980; Pieters, 1984; Houtman & Schinkelshoek, 1986) is over de toetsen van praktische vaardigheden veel minder bekend. De aandacht voor praktische vaardigheden en voor toepassingen daarvan in het onderwijs is van betrekkelijk recente datum (Sanders, 1980a en 1980b).

Ook binnen de medische opleidingen, het onderwerp van deze studie, neemt de belangstelling voor toetsen van praktische vaardigheden sterk toe. Na de introductie van praktijktoetsen in de vorm van Objective Structured Clinical Examinations in Engeland (OSCE's; Harden & Gleeson, 1979) en soortgelijke toetsen als 'Patient Based Instruments' (Stillman e.a., 1976) in de Verenigde Staten, hebben geleidelijk meer medische opleidingen dergelijke toetsvormen geïntroduceerd. De standaard vorm bestaat uit een circuit, dat wil zeggen een reeks van z.g. stations. Bij ieder station - een apart ingerichte ruimte met relevante hulpmiddelen en materiaal - wordt aan de student gevraagd een vaardigheid te tonen. Alle kandidaten starten gelijktijdig, ieder bij een verschillend station en na een vastgestelde periode (variërend tussen vijf en dertig minuten) gaan alle kandidaten naar een volgend station, totdat het gehele circuit is doorlopen. De kandidaten worden direct geobserveerd en beoordeeld door een aanwezige beoordelaar aan de hand van tevoren opgestelde gestandaardiseerde beoordelingsformulieren. De beoordelingsformulieren kunnen verschillende vormen hebben, variërend van globale oordelen door middel van rating scales, tot gedetailleerde checklists van het gewenste handelen.

Het doel van het toepassen van toetsen voor praktische vaardigheden kan nogal verschillen. Vaak wordt meer dan praktische vaardigheden alleen beoogd en worden ze gebruikt om het klinisch denken en handelen of probleemoplossend vermogen van studenten te meten. Hierbij krijgen zij vaak de vorm van van arts-patiënt simulaties. In alle gevallen is echter het direct observeren en zo objectief mogelijk registreren van het gedrag van kandidaten onder standaard instructies in standaard situaties kenmerkend.

Ondanks de toenemende belangstelling voor directe observatie als toetsvorm, is er slechts weinig bekend over de psychometrische eigenschappen ervan. Over toetsen voor praktische vaardigheden zoals toegepast in het lager en middelbaar beroepsonderwijs zijn spaarzame gegevens bekend (Sanders, 1980a en 1980b). Met betrekking tot toepassingen in medische opleidingen zijn er recentelijk enkele studies naar betrouwbaarheid en validiteit verricht.

Psychometrische bevindingen

Petrusa e.a. (1986) onderwierpen 99 studenten geneeskunde aan 17 stations met een totale toetstijd van ongeveer 2 uur en 15 minuten. Bij de afzonderlijke stations werden patiëntenproblemen aan de studenten voorgelegd door gebruikmaking van simulatiepatiënten. De beoordeling van de studenten werd door de simulatiepatiënten verricht aan de hand van betrekkelijk globale beoordelingslijsten. De bereikte betrouwbaarheidscoëfficiënt (alpha) bedroeg 0.57. Voor het

bereiken van een betrouwbaarheid van 0.80 - een waarde die wel als minimaal wordt beschouwd - zouden 51 stations nodig zijn met een totale toetsduur van ongeveer zeven uur. De interbeoordelaarsovereenstemming bleek gemiddeld aanvaardbaar (kappa 0.77). De correlaties, gecorrigeerd voor attenuatie, met stagebeoordelingen en een multiple choice test (NBME, nationale amerikaanse algemeen medische kennistoetsen) waren resp. 0.73 en 0.64. Het verband met kennistoetsen wordt als positief opgevat, omdat, naast het feit dat kennis wordt gezien als een noodzakelijke voorwaarde voor het goed kunnen beoefenen van praktische vaardigheden, de te meten concepten niet onafhankelijk van elkaar zijn: betere studenten zullen op de verschillende competenties beter zijn; de slechtere studenten scoren op elk gebied slechter. Afgezien van de benodigde toetslengte zijn dit aanvaardbare onderzoeksresultaten.

Newble & Swanson (ter perse), integreerden en analyseerden het toetsmateriaal van een aantal afnames van verschillende cohorten van in totaal 336 studenten geneeskunde. De schriftelijke stations buiten beschouwing latend, resulteerde een toets bestaande uit 5 stations met een totale gemiddelde toetsduur van een half uur in een generaliseerbaarheidscoëfficiënt van 0.31. Een aanvaardbare betrouwbaarheid zou pas bereikt worden bij ongeveer 6 uur toetsen. Beoordelingen vonden plaats door facultaire medewerkers en de correlaties met co-beoordelaars varieerden van middelmatig tot hoog (gemiddelde intraclass correlatie 0.75). Bij correlatie van de toets met een soortgelijke multiple choice toets werd een hoog verband gevonden (0.77; gecorrigeerd voor attenuatie).

Williams e.a. (1987), construeerden een zeer lange toets bestaande uit 18 stations van ongeveer 40 minuten per stuk (totale duur 15 uren). Deze bestond voor de ene helft uit een uitvoerend deel (anamnese en lichamelijk onderzoek) en de andere helft uit schriftelijke vragen over het betreffende patiëntenprobleem. De beoordelingen in het uitvoerende deel werden verricht door simulatiepatiënten die voor dit doel waren getraind. De toets werd afgenomen bij 70 studenten in hun examenjaar. De gerapporteerde betrouwbaarheidscoëfficiënt was 0.75. Er werd echter geen uitsplitsing gegeven naar uitvoerend en schriftelijk deel, noch werden beoordelaarsovereenstemmingen vermeld. Het verband met stagebeoordelingen was 0.65 en de correlaties met NBME part I en II resp. 0.53 en 0.51.

Stillman e.a. (1987) analyseerden een toets van 14 stations met een totale toetsduur van 3,5 uur. De opdrachten hadden betrekking op sociale vaardigheden (zoals interviewstijl en counseling) en beperkt fysisch diagnostisch onderzoek. Beoordelingen vonden plaats door simulatiepatiënten op basis van checklists en rating scales. De auteurs rapporteren een betrouwbaarheid van 0.78. De beoordelaarsovereenstemming bedroeg 0.66 voor de rating scales en 0.93 voor de checklists (Pearson correlaties). Correlatie met de NBME part I en II valt hier lager uit dan bij de andere studies met resp. 0.23 en 0.37; het verband met de stagebeoordelingen bleek 0.50 te zijn.

Hoewel er ook in Nederland initiatieven worden genomen voor de toetsing van medische vaardigheden (Hiemstra e.a., 1986; Van Rossum, 1985; Metz, 1986) zijn psychometrische gegevens zeer schaars voorhanden (Metz, 1984).

Met betrekking tot de aanwezige gegevens concludeert Swanson (1987), dat het zwakste punt van deze toetsvorm de toetslengte betreft. Ieder station kan opgevat worden als een enkel test-item en dientengevolge is al snel een groot

aantal stations noodzakelijk voor het verkrijgen van een voldoende generaliseerbaar resultaat. Dat maakt uiteraard de praktische uitvoerbaarheid van deze toetsvorm extra bezwaarlijk, te meer omdat er een zware logistieke belasting van de onderwijsorganisatie wordt gevergd. Interbeoordelaarsbetrouwbaarheid blijkt in het algemeen een geringer probleem dan inter-item (station) betrouwbaarheid. De validiteit (na correctie voor attenuatie) en de acceptabiliteit leiden er echter toe, dat de hier besproken toetsen over het algemeen gunstig worden beoordeeld.

De Maastrichtse vaardigheidstoets

Voor de onderwijskundige opzet van de Rijksuniversiteit Limburg - het probleemgestuurde leren (Schmidt, 1983) - is de integratie van theorie en praktijk een belangrijke doelstelling. Gepoogd wordt om praktische vaardigheden vanaf het begin van de studie onderdeel te laten zijn van de theoretische kennisvergaring. Daartoe is onder andere een onderwijskundige voorziening gecreëerd in de vorm van een vaardigheidslaboratorium, het zogenaamde Skillslab. Studenten kunnen in het Skillslab met alle mogelijke hulpmiddelen en voorzieningen in een veilige 'laboratorium-omgeving' medisch praktische vaardigheden aanleren en oefenen. De vaardigheden variëren van simpele motorische handelingen zoals het aanleggen van een verband of het meten van de bloeddruk, tot complexe fysisch diagnostische handelingen en interpretaties zoals die bijvoorbeeld bij het onderzoek van de thorax nodig zijn. Ook complexe sociale vaardigheden zoals het voeren van een slecht nieuws gesprek komen aan de orde. Gemiddeld besteedt een student 2 à 3 uur per week in het Skillslab. Voor een uitgebreide uiteenzetting van doelstellingen en uitwerkingen van het vaardigheidsonderwijs wordt verwezen naar Dochy & Van Luyk (1987).

Aan het einde van elk studiejaar wordt elke jaargroep (ongeveer 150 studenten) getoetst op hun vaardigheidsbeheersing door middel van een zogenaamde vaardigheidstoets. Ook in het vijfde en zesde studiejaar, wanneer klinische stages worden doorlopen, worden de studenten op deze wijze getoetst. De toets heeft consequenties voor de besluitvorming over de studievoortgang en vormt een geïntegreerd deel van het totale toetsingssysteem (cf. Verwijnen e.a., 1982).

Een vaardigheidstoets bestaat voor een individuele student uit een twee uren durend circuit van een variërend aantal stations (tussen de 6 en 11). Afhankelijk van de opdracht varieert de stationsduur van 10 tot 30 minuten. De keuze van de te toetsen vaardigheden is gebaseerd op een blauwdruk van 12 zogenaamde vaardigheidscategorieën. Afhankelijk van het genoten onderwijs wordt per categorie een vaardigheid geselecteerd en ondergebracht in een station. Bij een meerderheid van de stations wordt gebruik gemaakt van (simulatie)patiënten, die zo nodig speciaal getraind zijn voor de te spelen rol.

In elk station is tenminste één beoordelaar aanwezig, de z.g. observator. De observator is een deskundige op het gebied van de getoetste vaardigheid, en is tevoren getraind en voorbereid op zijn taak. In ongeveer 20 procent van de gevallen is ook een co-observator aanwezig, die onafhankelijk van de observator dezelfde beoordelingslijsten scoort. Voor het toetsen van één complete jaargroep van 150 studenten zijn twee werkdagen nodig.

Om het doorgeven van informatie over de inhoud van de toets te voorkomen wordt elke twee uur een ander, maar (op basis van de blauwdruk) inhoudelijk

parallel circuit ingericht. De observaties worden gescoord aan de hand van criterialijsten. Dit zijn gedetailleerde lijsten waarop de te toetsen vaardigheden zijn ontleed in operationele onderdelen. Een lijst kan variëren van 8 tot 90 items (gemiddeld ongeveer 30). De meeste criterialijst-items zijn dichotoom (uitgevoerd/niet uitgevoerd), maar bij sommige is een intermediaire codering mogelijk voor onvolledig uitgevoerde handelingen. Belangrijke items kunnen zwaarder worden gewogen. De score op een station wordt gevormd door de som van de itemscores; de score op een totaal circuit is de somscore van alle stationsscores en vormt de uitslag voor een individuele student. Scores worden uitgedrukt in percentages van het totaal mogelijk te behalen punten.

Resultaten op de vaardigheidstoets tellen mee in de beoordeling van de studievoortgang. Uitgangspunt bij de bepaling van de caesuur is een relatieve normering volgens de methode Wijnen (Wijnen, 1971).[1]

Enkele validiteitsstudies hebben aangetoond, dat de vaardigheidstoets (vooral in hogere jaargroepen) sterk correleert met kennismetingen (Van der Vleuten e.a., ter perse), predictieve waarde heeft ten opzichte van het functioneren in de klinische stages (Van der Vleuten e.a., 1987) en convergeert met oordelen van deskundigen op een aantal verschillende validiteitsaspecten (Van der Vleuten & Van Luyk, 1986). De onderhavige studie zal zich beperken tot de betrouwbaarheid van de vaardigheidstoets.

Vraagstelling en verwachtingen

De stabiliteit van de toets zal worden onderzocht door analyse van een drietal variatiebronnen: verschillen tussen beoordelaars, verschillen in moeilijkheidsgraad tussen vaardigheden (stations) en verschillen in moeilijkheidsgraad tussen parallelle circuits binnen een toets. Een vierde variatiebron, verschillen ontstaan door gebruik van verschillende (simulatie)patiënten met eenzelfde rol, zal hier niet nader worden onderzocht omdat de benodigde gegevens (vooralsnog) ontbreken. Uit recentelijk onderzoek blijkt, dat weliswaar op stationsniveau statistisch significante verschillen tussen simulatiepatiënten worden gevonden (Hiemstra e.a., 1987; Dawson-Saunders e.a., 1987), maar dat de variatie over de gehele test heen aanvaardbaar klein is (Swanson & Norcini, in voorbereiding). Tot slot betreft een mogelijke vijfde variatiebron de stabiliteit van de afzonderlijke stations. Het ligt voor de hand dat stations met korte beoordelingslijsten minder stabiele resultaten opleveren dan beoordelingen op grond van langere lijsten. Dit kan weer repercussies hebben voor de totale toetsscore. Psychometrisch is dit verband echter nog niet eenduidig te beschrijven (Van der Vleuten, 1987) en nader onderzoek dient eerst nog plaats te vinden.

Gegeven de bevindingen uit de studies die in de inleiding werden genoemd is te verwachten, dat ook hier het aantal te toetsen vaardigheden (stations) voor het verkrijgen van een betrouwbaar resultaat een probleem zal zijn. Voortvloeiend uit de te verwachten verschillen in moeilijkheidsgraad van stations, zullen moeilijkheidsgraadverschillen tussen parallelle circuits wellicht ook optreden.

---

[1]Recentelijk zijn aan de regel absolute grenzen toegevoegd, waarboven of beneden altijd een (on)voldoende wordt behaald.

De beoordelaarsovereenstemming zou voor de vaardigheidstoets gunstig moeten uitvallen, te meer daar gewerkt wordt met getrainde en deskundige beoordelaars, werkend met gedetailleerde operationele criterialijsten.

Methode

Subjecten

In totaal waren 18 toetsen (3 academische jaren x 6 toetsen) voor de analyse beschikbaar. Binnen iedere toets kan er sprake zijn van een verschillend aantal circuits, terwijl ieder circuit een verschillend aantal stations en een verschillend aantal studenten kan bevatten. Tabel 1 bevat hiervan een overzicht.

Tabel 1: Beschrijving van vaardigheidstoetsen over de periode 1984-1987.

Academisch jaar	Jaar	Aantal circuits	Aantal stations	Totale toetsduur in min.	Aantal studenten
84/85	1	5	3	30	143
	1	3	4	60	141
	2	4	9	120	141
	3	3	10	120	126
	4	9	6/7	120	116
	5	4	7	120	82
	6	4	8/9	120	51
85/86	1	4	3	30	149
	1	6	4	60	151
	2	3	9	120	121
	3	4	10	120	140
	4	4	6/7	120	128
	5	3	7	120	110
	6	4	8/9	120	83
86/87	1	6	3	30	150
	1	4	4	90	153
	2	4	9	120	143
	3	3	10	120	125
	4	3	6/7	120	135
	5	3	8	120	117
	6	2	8/9	120	47
	6	2	8	120	55

De toetsen in het eerste jaar bestonden uit twee delen die op afzonderlijke tijdstippen werden afgenomen en waaraan alle eerstejaars deelnamen, maar waarover een gecombineerde uitslag werd gegeven. In 1986/1987 werd de zesdejaars-toets eveneens twee maal afgenomen. In dit geval echter namen studenten deel aan hetzij de ene hetzij de andere afname, afhankelijk van hun vorderingen in de stages.

Sommige toetsen bestaan uit nogal veel circuits met weinig studenten (minder dan 20). Soms bleek dit grote aantal circuits noodzakelijk in verband met

logistieke en organisatorische problemen (bijvoorbeeld uitval van apparaten of fantomen).

Analyseprocedure beoordelaarsovereenkomst

In totaal waren 3402 paren observator/co-observator scores beschikbaar voor analyse. Deze werden op twee manieren geanalyseerd.

Op stationsscore-niveau werden de gepaarde scores van observator/co-observator geanalyseerd met een 'random effects beoordelaars genest binnen stations' variantie-analyse (ANOVA), geen rekening houdend met student-identiteit: het i:p design in termen van de generaliseerbaarheidstheorie (Cronbach e.a., 1982; Brennan, 1983). Variantiecomponenten werden vervolgens geschat op basis van de ANOVA mean squares volgens standaard procedures (Brennan, 1983) en een generaliseerbaarheidscoëfficiënt werd bepaald (voor één beoordeling), waarbij de gemiddelde verschillen tussen observatoren in de error term werden betrokken.

Op het itemniveau van de criterialijsten binnen stations werd een percentage overeenkomst tussen observator en co-observator scores berekend. Discrepanties tussen item-overeenkomst en stationsscore-overeenkomst kunnen kwalitatieve aanduidingen geven over de aard van de beoordeling.

Analyseprocedure reproduceerbaarheid van scores en beslissingen

De beoordelingen per getoetste vaardigheid, d.w.z. de stationsscores, worden opgevat als items en vormen de basis voor de betrouwbaarheidsschatting van de totale toets. Voor ieder circuit binnen een jaargroep werd een 'random effects personen x stations' ANOVA uitgevoerd (p x i design) en werden variantiecomponenten geschat voor personen, stations en error (voor de gesplitse eerstejaarstoetsen werden de variantiecomponenten van beide delen gesommeerd). De variantiecomponenten werden vervolgens gemiddeld per en over jaargroepen (gewogen naar steekproefgrootte) om een schatting van de totale effecten te verkrijgen.

De berekening van de betrouwbaarheidscoëfficiënten zal op verschillende wijze worden uitgevoerd, ter illustratie van verschillende mogelijke interpretaties. Uitgaande van een norm-georiënteerd standpunt zullen generaliseerbaarheidscoëfficiënten worden berekend, waarbij geen rekening wordt gehouden met moeilijkheidsgraadverschillen tussen stations. Daarnaast zal een berekening plaatsvinden volgens een domein-georiënteerd perspectief, enerzijds door verschillen in moeilijkheidsgraad in de error term op te nemen (dependability coefficient) en anderzijds rekening houdend met verschillende zak/slaag grenzen (adjusted phi coefficient; Brennan & Kane, 1977). Feitelijke berekeningen vonden plaats met het programma GENOVA (Crick & Brennan, 1983).

Resultaten

Beoordelaarsovereenstemming

Tabel 2 bevat beoordelaarsovereenstemmingen uitgesplitst naar jaargroep.

Tabel 2: Beoordelaarsovereenstemmingen naar jaargroep en totaal.

Jaar	Aantal paren	Intraclass correlatie	Percentage overeenkomst
1	581	0.85	87
2	682	0.86	84
3	699	0.84	84
4	527	0.86	85
5	443	0.75	84
6	470	0.82	87
1-6	3402	0.84	85

De intraclass correlaties zijn redelijk hoog en vertonen weinig verschillen tussen jaargroepen. De waarde in jaar 5 vormt hierop een uitzondering. Wellicht kan dit worden toegeschreven aan de geringere ware score variantie in jaar 5 (zie variantiecomponenten tabel 4). Het percentage overeenstemming op itemniveau van de criterialijsten van jaar vijf is overeenkomstig de overige jaren. Voor het overige laten de overeenstemmingspercentages eenzelfde beeld zien. Hoewel zij niet corrigeren voor toeval en gevoelig zijn voor de frekwenties in de randtotalen, blijken de percentages een betrekkelijk aanvaardbaar en stabiel resultaat over de verschillende jaren heen op te leveren. Concluderend kan worden gesteld, dat de totale error die kan worden toegeschreven aan beoordelaarsfouten laag is. Op het niveau van de totale toets zullen de verschillen een nog kleinere rol spelen, omdat iedere student door een meervoud van observatoren wordt beoordeeld en aangenomen mag worden dat beoordelaarsverschillen zullen uitmiddelen.

Een tweede analyse van beoordelaarsovereenstemming betreft uitsplitsing naar de verschillende inhoudsgebieden. Het ligt voor de hand dat sommige vaardigheden moeilijker te beoordelen zijn dan andere. Tabel 3 bevat een uitsplitsing naar (blauwdruk)categorieën van de vaardigheidstoets over alle jaren heen.

Voor de categorie sociale vaardigheden blijken beide overeenkomstmaten laag. Gegeven de subjectiviteit in de beoordeling van sociale vaardigheden is dit geen opmerkelijk resultaat. De criterialijsten die gebruikt worden bevatten allerlei proces- en uitkomstenmaten, die interpretaties van de beoordelaar vergen. Ondanks de zorgvuldige samenstelling van de beoordelingslijsten voor sociale vaardigheden (cf. Kraan & Crijnen, 1987) en de specifieke voorbereidende trainingen aan de hand van een uitgebreide handleiding, blijft de beoordeling een meer subjectieve aangelegenheid, leidend tot verschillen tussen beoordelaars. De discrepantie voor 'receptuur' tussen de relatief lage intraclass correlatiecoëfficiënt en het hoge overeenkomstpercentage kan worden verklaard

uit de eenvoud van deze vaardigheden en het geringe aantal items op de criterialijst. Als gevolg daarvan wordt algemeen zeer hoog gescoord resulterend in een geringe ware score variantie, terwijl anderzijds de overeenkomst tussen beoordelaars op itemniveau hoog is.

Tabel 3: Beoordelaarsovereenstemmingen uitgesplitst naar vaardigheidscategorieën over alle jaargroepen.

Categorie	Aantal paren	Intraclass correlatie	Percentage overeenkomst
Neurologische vaardigheden	122	0.90	88
Therapeutische vaardigheden	476	0.84	89
Abdomen	219	0.82	89
Thorax/perifere circulatie	306	0.89	85
Oogheelkundige vaardigheden	168	0.73	87
Gynaecol./Obst. vaardigheden	216	0.87	82
Bewegingsapparaat	245	0.86	85
KNO vaardigheden	139	0.77	87
Peadiatrische vaardigheden	41	0.82	87
Sociale vaardigheden	511	0.66	75
Laboratorium vaardigheden	799	0.84	87
Receptuur vaardigheden	45	0.71	94
Diversen[1]	115	0.84	83

[1] Experimentele stations met combinaties tussen verschillende categorieën

Variantiecomponenten

Tabel 4 bevat de geschatte variantiecomponenten resulterend uit de generaliseerbaarheidsanalyses uitgesplitst per jaargroep en over de totale data-set.

Tabel 4: Geschatte variantiecomponenten per jaargroep en over alle jaargroepen.

	Geschatte variantiecomponenten			Percentage van totale variantie		
Jaar	Personen	Stations	Error	Personen	Stations	Error
1	67.1406	208.2347	284.5977	11.99	37.19	50.82
2	68.2282	81.7080	212.3640	18.83	22.55	58.62
3	47.0295	159.2826	190.5516	11.85	40.14	48.01
4	51.7329	72.5922	140.8575	19.51	27.37	53.12
5	18.5279	89.0927	150.5470	7.18	34.51	58.31
6	68.1309	92.0852	151.9589	21.82	29.50	48.68
1-6	54.1622	122.0340	195.0798	14.59	32.87	52.54

Zoals in het algemeen bij toetsen gebruikelijk is bestaat ook hier de grootste variantiecomponent uit de error. Gerelateerd aan de variantiecomponent van personen is de variantiecomponent voor stations eveneens aanzienlijk. Blijkbaar

zijn er aanmerkelijke verschillen in de moeilijkheidsgraad van de verschillende stations.

Bij vergelijking van de variantiecomponenten over de verschillende jaren valt de lage persoonscomponent in jaar 5 op. Het vijfde studiejaar is het eerste stagejaar en de studenten doorlopen allerlei verschillende klinische programma's. In de keuze van de chronologie van de stages bestaat enige vrijheid. Hierdoor zouden verschillen in achtergrond kunnen bestaan die een verklaring kunnen geven voor de lage persoonscomponent. In geringe mate blijkt dat ook uit de toename van de error component, waarin de persoon x station interactie term besloten zit. Een tweede verklaring ligt mogelijk in de verschillen in zwaarte van de verschillende toetsen. Beslissingen over de studievoortgang vinden plaats volgens de twee-fasenstructuur in jaren 1, 4 en 6. In de besluitvorming hebben vaardigheidstoetsen in de overige jaren een minder zwaar gewicht. De gemiddeld behaalde scores in de tussenliggende jaren liggen dan ook doorgaans iets lager. Voor jaar vijf geldt daarenboven dat er nauwelijks ruimte is voor voorbereiding op de toets in verband met drukte van de stage-routine.

Een mogelijke invloed van het feit, dat er beslissingen worden genomen over de studievoortgang is ook terug te vinden in de jaren 4 en 6 waar de persoonscomponenten relatief hoog zijn, maar minder in het eerste jaar. Deze lage persoonscomponent lijkt ook niet verklaard te kunnen worden door de betrekkelijke eenvoud van de vaardigheden in het eerste jaar. Gecombineerd met een goede voorbereiding zou eenvoud van vaardigheden kunnen resulteren in algemeen hoge scores, leidend tot een lage persoonscomponent, maar ook in een lage stationscomponent. In feite is deze laatste echter een van de grootste van alle jaargroepen. Deze bevinding zou erop kunnen wijzen, dat er in het eerste jaar één of enkele stations zitten, die als uitbijter functioneren. De alternatieve verklaring kan natuurlijk ook gelegen zijn in de kwaliteit van de gebruikte beoordelingslijsten in het eerste jaar. Nadere (inhoudelijke) analyse zal dit moeten uitwijzen.

Samenvattend kan gesteld worden dat de variantiebronnen behorend bij stations en error relatief groot zijn, waardoor noodzakelijkerwijs het aantal te toetsen vaardigheden eveneens groot zal moeten zijn voor het verkrijgen van een reproduceerbaar resultaat. Het betekent ook dat parallelle toetsvormen, zoals die bijvoorbeeld noodzakelijk zijn in de Maastrichtse situatie, zullen variëren in moeilijkheid, waardoor bij een absolute interpretatie sommige studenten benadeeld kunnen worden. De reproduceerbaarheid van scores zal worden geanalyseerd in de navolgende paragraaf.

Reproduceerbaarheid van scores

De schattingen van de variantiecomponenten uit tabel 4 vormden de basis voor de schatting van de tussen-stations-betrouwbaarheid en standaardmeetfout, geprojecteerd naar verschillende testlengtes en score-interpretaties. De berekeningen zijn alleen gebaseerd op de gecombineerde gewogen gemiddelde resultaten over jaar 1 tot en met 6. Tabel 5 bevat een overzicht.

Uitgaande van een toetsbenadering waarin alleen de relatieve posities van personen van belang zijn en niet het absolute scoreniveau, is de generaliseerbaarheidscoëfficiënt de aangewezen index. Deze is (hier) vergelijkbaar met Cronbach's alpha. Het is duidelijk dat nogal wat tijd wordt vereist voor het

verkrijgen van een adequate betrouwbaarheid. Pas na ongeveer drie uur wordt een minimaal aanvaardbaar niveau bereikt.

Tabel 5: Reproduceerbaarheid van scores bij verschillende test-lengtes voor norm-georiënteerde interpretatie (generaliseerbaarheidscoëfficiënten: G) en domein-georiënteerde interpretatie (dependability coëfficiënten: D) en bijbehorende standaardmeetfout (SEM).

Toetstijd in uren	Aantal stations	Norm-georiënteerde interpretatie		Domein-georiënteerde interpretatie	
		G	SEM	D	SEM
1	4	0.53	6.98	0.41	8.90
2	8	0.69	4.94	0.58	6.30
3	12	0.77	4.03	0.67	5.14
4	16	0.81	3.49	0.73	4.45
5	20	0.85	3.12	0.77	3.98
10	40	0.92	2.21	0.87	2.82

In het domein-georiënteerde perspectief wordt aan de scores een absolute betekenis gehecht en dienen moeilijkheidsgraadverschillen als error te worden opgevat. Uit tabel 5 blijkt dat in dat geval meer dan vijf uur toetstijd noodzakelijk wordt.

De standaardmeetfouten zijn weergegeven in de oorspronkelijke score-schaal (percentages). Bij een reproduceerbaarheidscoëfficiënt van ongeveer 0.80 is de standaardmeetfout nog altijd van respectabele grootte (ongeveer 4 procent).

Een alternatieve domein-georiënteerde benadering betreft de verschuiving van de aandacht van de reproduceerbaarheid van scores naar de reproduceerbaarheid van zak/slaag beslissingen. Bij examens is men geïnteresseerd in het scheiden van voldoende en onvoldoende prestaties, vaak meer dan in de mate van niveaubeheersing. Dat kan, afhankelijk van de caesuur en het (gemiddeld) behaalde examenresultaat, een heel ander beeld opleveren van de reproduceerbaarheid. De gemiddelde score behaald over alle vaardigheidstoetsen was 77.09 procent. Tabel 6 bevat betrouwbaarheden berekend voor daarbijbehorende verschillende zak/slaag grenzen.

Naarmate de caesuur verder verwijderd ligt van de gemiddeld behaalde score verandert de betrouwbaarheid van de beslissing aanzienlijk. Afhankelijk van de toetssituatie (bijvoorbeeld de mate van voorbereiding) en het vereiste niveau kunnen daarmee nogal grote verschillen in de benodigde toetstijd ontstaan. In dit geval levert een wat lager vereist niveau van 60 procent slechts één uur benodigde toetstijd op, hetgeen aanmerkelijk korter is dan bij de hierboven gevonden betrouwbaarheidscoëfficiënten. Voor een norm dichter bij het behaalde gemiddelde zijn aanzienlijk langere toetstijden nodig. Overigens moet er rekening mee worden gehouden, dat de zak/slaag grens en gemiddeld behaalde score niet onafhankelijk zijn van elkaar. Het is denkbaar dat de gemiddeld behaalde score zich aanpast aan de caesuur.

Tabel 6: Reproduceerbaarheid (adjusted phi coëfficiënten) bij verschillende zak/slaag percentages.

Toetstijd in uren	Aantal stations	Zak/slaag percentage[1] 60%	70%	80%	90%
1	4	0.80	0.48	0.29	0.71
2	8	0.89	0.69	0.54	0.84
3	12	0.93	0.78	0.66	0.89
4	16	0.94	0.83	0.74	0.91
5	20	0.96	0.86	0.78	0.93
10	40	0.98	0.93	0.88	0.96

[1] Behaalde gemiddelde score 77.09%

Circuitverschillen

De circuits binnen één vaardigheidstoets van één jaargroep kunnen worden opgevat als paralleltoetsen. Om benadeling van studenten te voorkomen zouden verschillende circuits van gelijke moeilijkheid behoren te zijn. Door verschillen in moeilijkheidsgraad in aanmerking te nemen bij de berekening van de betrouwbaarheid, wordt reeds een schatting gegeven van de betrouwbaarheid volgens een norm-gerichte interpretatie voor een situatie waarin niet iedereen dezelfde items krijgt aangeboden. De D-coëfficiënten uit tabel 5 zijn aldus berekend en vormen in dat geval de juiste norm-georiënteerde interpretatie voor de betrouwbaarheid. De studenten krijgen in de vaardigheidstoets echter geen willekeurige set van stations voorgelegd, maar zij worden groepsgewijs aan dezelfde stations onderworpen. Directe vergelijking van de grootte van het effect van circuitverschillen kan worden bekeken door variaties in circuitgemiddelden te vergelijken.

Studenten worden random toegewezen aan de verschillende circuits. Toetsing van de verschillen vond plaats aan de hand van een one-way variantie-analyse over de scores van studenten, met circuitnummer als groepsvariabele. Tabel 7 bevat de resultaten.

Van de vijf statistisch significante verschillen worden er drie in het eerste jaar geconstateerd. Hier is echter sprake van korte deeltoetsen met een gering aantal stations (3 of 4). Het verschil in jaar 4 van 1984/1985 is evenmin erg representatief, aangezien hier sprake was van een (noodgedwongen) opdeling in veel circuits met weinig personen. Een en ander wijst erop dat de circuits niet al te sterk van elkaar verschillen in moeilijkheidsgraad. In de praktijk hebben deze gegevens nooit aanleiding gegeven tot correctie van individuele scores van studenten.

Tabel 7: Variantie-analyses over circuitgemiddeldes binnen een toets.

Ac. jaar	jr	Gemiddeld behaalde procentuele score op circuit 1	2	3	4	5	6	7	8	9	F	p
84/85	1	80	80	80	82	81					0.119	0.98
	1	69	79	80							16.485	0.00*
	2	72	76	74	73						1.060	0.37
	3	64	66	68							2.344	0.10
	4	76	66	71	73	72	71	72	78	73	2.296	0.03*
	5	76	76	79	79						1.034	0.38
	6	72	75	70	71						1.487	0.23
85/86	1	80	88	84	85						3.918	0.01*
	1	80	84	83	80	81	78				0.585	0.71
	2	71	71	77							5.291	0.01*
	3	74	74	76	74						0.694	0.56
	4	79	83	81	81						1.453	0.23
	5	71	70	73							1.811	0.17
	6	82	86	82	81						1.229	0.30
86/87	1	79	79	77	84	75	77				1.101	0.41
	1	80	84	78	81						3.792	0.01*
	2	76	78	78	69						2.072	0.11
	3	73	76	74							2.767	0.07
	4	83	82	83							0.258	0.77
	5	77	76	77							0.685	0.51
	6	80	79								0.192	0.66
	6	83	85								0.411	0.52

* $p < .05$

Discussie

De resultaten uit deze studie komen sterk overeen met de bevindingen uit de literatuur. De verschillen tussen beoordelaars zijn niet erg groot en de resulterende betrouwbaarheden zijn aanvaardbaar. Bovendien blijkt dit betrekkelijk invariant over de verschillende jaargroepen. Wel kwam naar voren dat de verschillen tussen beoordelaars varieerden met de inhoud van de vaardigheden. Met name bij de beoordeling van sociale vaardigheden bleek de overeenkomst relatief laag. Geconcludeerd kan echter worden dat beoordelaarsverschillen geen ernstige bedreiging vormen voor de toetsing van praktische vaardigheden, te meer niet wanneer elke kandidaat door verschillende beoordelaars wordt gezien, waardoor beoordelaarskenmerken elkaar zullen neutraliseren door uitmiddeling.

Veel groter en belemmerender is de variatiebron *tussen* de verschillende vaardigheden. Conform de bevindingen uit de literatuur bleken de stations nogal in moeilijkheidsgraad te verschillen. De bijbehorende variantiecomponent bleek relatief groot in vergelijking tot de variatie tussen personen. De algemene error variantie bleek (zoals gewoonlijk) de grootste foutenbron. In verhouding

tot de geobserveerde variantie is daarmee de ware score variantie betrekkelijk laag. Om tot een adequate totale betrouwbaarheid te komen dient als gevolg daarvan het aantal te toetsen vaardigheden groot te zijn waardoor lange toetstijden nodig worden.

Dit werd duidelijk door de uitkomsten van de D-studies waarin de betrouwbaarheid werd geprojecteerd naar verschillende toetslengtes. Een adequate betrouwbaarheid bleek pas verkregen te kunnen worden na tenminste enkele uren toetstijd, afhankelijk van de interpretatie en het perspectief voor het gebruik van de toets. Bij een norm-georiënteerde benadering is men slechts geïnteresseerd in de relatieve posities van de kandidaten en worden hogere betrouwbaarheden geboekt dan bij de domein-georiënteerde benadering. Bij deze laatste is men geïnteresseerd in het absoluut behaalde niveau en worden verschillen in moeilijkheidsgraad tussen de verschillende vaardigheden als meetfout beschouwd, resulterend in lagere betrouwbaarheidsschattingen. Voor een norm-georiënteerde interpretatie was hier tenminste drie uren toetstijd noodzakelijk, terwijl voor de domein-georiënteerde interpretatie minimaal vijf uren nodig bleek. De standaardmeetfout lag in dit geval rond de 4 procent van de oorspronkelijke schaalscores.

De in de literatuur vermelde betrouwbaarheidscoëfficiënten zijn allen berekend volgens een norm-georiënteerde benadering. De betrouwbaarheidscoëfficiënten uit het hier gerapporteerde onderzoek blijken iets gunstiger te zijn dan in de literatuur vermelde resultaten. Meerdere verklaringen zijn hiervoor mogelijk. De meest waarschijnlijke verklaring is wellicht de invloed van de inhoud van de toets. Zoals aangegeven beogen de meeste toetsen uit de vermelde studies een bredere competentie dan de vaardigheidstoets in het kader van de Maastrichtse situatie. Deze beperkt zich meer tot medisch praktische vaardigheden, terwijl men in de andere studies vaker (hogere) cognitieve vaardigheden tracht te meten, zoals probleemoplossen en kennis. Beperking tot toetsing van praktische vaardigheden betekent dat een kleiner domein wordt beoogd, waardoor het generalisatiedomein makkelijker wordt gedekt bij steekproeftrekking en samenstelling van de toets. Een tweede verklaring ligt mogelijk in de inmiddels opgebouwde routine van vaardigheidstoetsing. Kenmerkend voor de Maastrichtse situatie is de grootschalige werkwijze, welke in geen van de andere studies voorkomt. De samenstelling, organisatie en afname is routine geworden, zowel docenten als studenten weten wat van hen verwacht wordt en wellicht dat deze exogene kenmerken de psychometrische kwaliteit van de vaardigheidstoets ten goede komen.

Ter voorkoming van lange toetstijden is recentelijk gesuggereerd om de aandacht te verschuiven naar de kwaliteit van de genomen beslissing (Swanson & Norcini, in voorbereiding) in plaats van naar de kwaliteit van de scores. In praktijksituaties waar de toetsen gehanteerd worden in de besluitvorming is dit een zinvolle strategie. In deze studie zijn betrouwbaarheidscoëfficiënten berekend voor enkele verschillende zak/slaag grenzen en werden aanmerkelijke verschillen in betrouwbaarheid en benodigde toetslengte gevonden. Naarmate de caesuur verder afligt van het algemene gemiddelde neemt de betrouwbaarheid van de beslissing sterk toe en kan worden volstaan met aanmerkelijk kortere toetstijden. Hierbij valt echter aan te tekenen dat de zak/slaag grens en de toetsprestatie niet onafhankelijk zijn van elkaar. In de praktijk blijken prestaties

van studenten zich vaak aan te passen aan het gewenste niveau voor een voldoende beoordeling.

Samenvattend kan echter geconcludeerd worden dat de toetslengte van observatietoetsen voor praktische vaardigheden een zwakke schakel vormt voor de betrouwbaarheid. Op zich is dat een wat ontmoedigend gegeven, zeker als de logistieke belasting van dit soort toetsen in aanmerking wordt genomen. Voor toetsconstructeurs betekent dit dat zuinigheid betracht moet worden bij de samenstelling. Er zal extra nauwkeurig op gelet moeten worden dat geen vaardigheden worden getoetst die ook, maar vooral ook makkelijker, met andere meetinstrumenten gemeten kunnen worden.

Uiteraard bestaat de rechtvaardiging van observatietoetsen niet uitsluitend uit de kwaliteit van de psychometrische kenmerken. Vanuit een onderwijskundig standpunt bezien kunnen deze toetsen grote invloed hebben op het onderwijs zelf (Newble & Jaeger, 1983; Stillman & Swanson, 1987). Toetsen voor praktische vaardigheden stimuleren het aanleren van vaardigheden, kunnen motiverend werken voor docenten en kunnen aanleiding geven tot standaardisatie (consensus) over wat vaardigheden zijn en hoe ze moeten worden aangeleerd (Bouhuijs e.a., 1987).

Niettemin zou de techniek aan gewicht winnen als ook voldoende betrouwbare informatie zou kunnen worden verkregen. Daarbij zijn enkele mogelijkheden denkbaar.

In de eerste plaats zou men lange toetsen kunnen gebruiken, zij het dat dit een zware logistieke inspanning zal vergen. Wel kan in dit verband worden opgemerkt, dat het uitbreiden van de toets met meer stations een gunstiger effect zal hebben op de betrouwbaarheid dan het inzetten van meerdere beoordelaars bij dezelfde stations. Aangezien de variatie tussen vaardigheden groter is dan de variatie tussen beoordelaars, zal het verlengen van de toets de betrouwbaarheid sterker beïnvloeden dan het inzetten van meer beoordelaars.

In de tweede plaats kan worden aangeraden om toetsen van praktische vaardigheden te combineren met schriftelijke toetsen (Swanson, 1987). De efficiëntie van de schriftelijke toetsvorm kan worden gecombineerd met de adequate predictieve waarde ervan (Van der Vleuten e.a., ter perse). De gecombineerde score zal een betrouwbaarder resultaat opleveren.

Een derde optie betreft de hiervoor genoemde verschuiving van reproduceerbaarheid van toetsscores naar de reproduceerbaarheid van zak/slaag beslissingen. Wellicht zou dit kunnen worden gecombineerd met een sequentiële testprocedure, waarin per individu onderscheid wordt gemaakt in toetstijd. Bij hoger scorende kandidaten kan worden volstaan met een korte toets, terwijl bij lager scorende studenten langere toetstijden nodig zullen zijn.

Tot slot is een vierde mogelijkheid, voor zover inhoudelijk mogelijk, het verkorten van de stationsduur en de gewonnen tijd te compenseren door meer vaardigheden te toetsen. Voorlopige aanwijzingen duiden op een per saldo grotere winst in betrouwbaarheid door het verkorten van de stationslengte, zelfs wanneer dat resulteert in een sterke reductie van de precisie van de meting binnen een enkel station (Van der Vleuten, 1987).

Bezien zal moeten worden of de conclusies uit deze studie uitsluitend beperkt zijn tot praktische vaardigheden in het medische domein. Voor meer simpele praktische vaardigheden, of meer pure psychomotorische vaardigheden

zoals bijvoorbeeld technische vaardigheden uit het lager beroepsonderwijs, kan een hele andere situatie gelden. Verder psychometrisch onderzoek zal dat moeten uitwijzen. Wat betreft meer eenvoudige handelingsgerichte verpleegkundige vaardigheden hebben de eerste resultaten niettemin soortgelijke conclusies opgeleverd (Van der Vleuten e.a., in voorbereiding).

Referenties

Bouhuijs, P.A.J., Vleuten, C.P.M. van der & Luyk, S.J. van (1987) The OSCE as a part of a systematic skillstraining approach. *Medical Teacher*, 9, 183-191.

Brennan, R.L. (1983) *Elements of Generalizability Theory*. Iowa: American College Testing Program.

Brennan, R.L. & Kane, M.T. (1977) An index of dependability for mastery tests. *Journal of Educational Measurement*, 14, 277-289.

Crick, J.E. & Brennan, R.L. (1983) *Manual for GENOVA: A Generalized Analysis of Variance System*. Iowa: American College Testing Program.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Dawson-Saunders, B., Verhulst, S.J., Marcy, M. & Steward, D.E. (1987) Variability in standardized patients and its effect on student performance. In: I.R. Hart & R.M. Harden, (Eds.), *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.

Dochy, F.J.R.C. & Luyk, S.J. van (Eds.) (1987) *Handboek Vaardigheidsonderwijs*. Lisse: Swetz & Zeitlinger.

Harden, R.M. & Gleeson, F.A. (1979) ASME Medical Education Booklet No. 8. Assessment of clinical competence using an objective structured clinical examination (OSCE), *Medical Education*.

Hiemstra, R.J., Bender, W., Scherpbier, A., Lunsen, H.W., Vries, P.G.M. de & Soeters, D. (1986) An objective structured clinical examination (OSCE) in Groningen, The Netherlands. In: I.R. Hart, R.M. Harden, R.M. & J.H. Walton, (Eds.), *Newer Developments in Assessing Clinical Competence*. Montreal, Heal Publications.

Hiemstra, R.J., Scherpbier, A.J.J.A. & Roze, B.J. (1987) Assessing history-taking skills or ... simulated patients' peculiarities. In: I.R. Hart & R.M. Harden, (Eds.), *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.

Houtman, I. & Schinkelshoek, D. (1986) *Toetsen van praktische vaardigheden*. IFLO docentenopleiding en Afdeling Onderwijs Research, Vrije Universiteit, Amsterdam.

Klerk, L.F.W. de (1980) *Het leren van psychomotorische vaardigheden: een onderwijspsychologische benadering*. Deventer: Van Loghum Slaterus.

Kraan, H. & Crijnen, A. (1987) *The Maastricht History Taking and Advice Checklist: Studies of instrumental utility*. Academisch proefschrift Rijksuniversiteit Limburg.

Metz, J.C.M. (1984) *Medische competentie, Een onderzoek naar de betrouwbaarheid en de validiteit van het Gestructureerd Klinisch Examen*. Dissertatie, Nijmegen.

Metz, J.C.M. (1986) Het gestructureerd klinisch examen. *Nederlands Tijdschrift voor de Geneeskunde*, 130, 2091-2093.

Newble, D.I. & Jaeger, K. (1983) The effect of assessments and examinations on the learning of medical students. *Medical Education*, 17, 165-171.

Newble, D.I. & Swanson, D.B. (Ter perse) Psychometric characteristics of the Objective Structured Clinical Examination. *Medical Education.*

Petrusa, E.R., Blackwell, T., Parcel, S. & Saydjari, C. (1986) Psychometric properties of the Objective Clinical Exam as an instrument for final evaluation. In: I.R. Hart, R.M. Harden, R.M. & J.H. Walton, (Eds.), *Newer Developments in Assessing Clinical Competence*, Montreal: Heal Publications.

Pieters, J.M. (1984) Praktische vaardigheden in het voortgezet onderwijs. In: L.F.W. De Klerk & A.M.P. Knoers, (Eds.), *Onderwijspsychologisch Onderzoek.* Lisse: Swets & Zeitlinger.

Rossum, H.J.M. van (1985) Honderd groepen Alco. In: Th. W.M. Hoeks, E. van der Putte & H.J.M. van Rossum, (Eds.), *Vijf jaar Alcoschap, een geslaagd experiment.* Faculteit der Geneeskunde, Leiden.

Sanders, P.F. (1980a) Een inleiding tot toetsen praktische vaardigheden. *Specialistisch Bulletin, nr. 5*, Cito Arnhem.

Sanders, P.F. (1980b) Een procedure voor de constructie toetsen praktische vaardigheden. *Specialistisch Bulletin, nr. 9*, Cito Arnhem.

Schmidt, H.G. (1983) Problem based learning: Rational and description. *Medical Education, 17*, 11-16.

Stillman, P.L., Sabers, D. & Redfield, D. (1976) The use of paraprofessionals to teach and evaluate interviewing skills in medical students. *Pediatrics, 57*, 769-774.

Stillman, P.L. & Swanson, D.B. (1987) Ensuring the clinical competence of medical school graduates through standardized patients. *Archives of Internal Medicine, 147*, 1049-1052.

Stillman, P.L., Regan, M.B. & Swanson, D.B. (1987) A diagnostic fourth year performance assessment. *Archives of Internal Medicine, 105*, 762-771.

Swanson, D.B. (1987) A measurement framework for performance based tests. In: I.R. Hart & R.M. Harden (Eds.), *Further Developments In Assessing Clinical Competence.* Montreal: Can-Heal.

Swanson, D.B. & Norcini, J.J. (In voorbereiding) Factors influencing the reproducibility of tests using standardized patients.

Verwijnen, G.M., Imbos, Tj., Snellen, H., Stalenhoef, B., Pollemans, M., Luyk, S. van, Sprooten, M., Leeuwen, Y. van & Vleuten, C.P.M. van der (1982) The evaluation system of the medical school of Maastricht. *Assessment and Evaluation in Higher Education, 3*, 225-244.

Vleuten, C.P.M., van der (1987) Intracase versus intercase reliability: A trade-off? *PES-Publ. nr. 178*, Interne publicatie Rijksuniversiteit Limburg.

Vleuten, C.P.M. van der & Luyk, S.J. van (1986) A validity study of a test for clinical and technical medical skills. In: I.R. Hart, Harden, R.M. & Walton, H.J. (Eds.), *Newer Developments in Assessing Clinical Competence.* Montreal: Heal Publications.

Vleuten, C.P.M. van der, Luyk, S.J. van & l'Espoir, N.E.J.C. (1987) Effecten van vaardigheidsonderwijs. In: F.J.R.C. Dochy & S.J. van Luyk, (Eds.), *Handboek Vaardigheidsonderwijs*. Lisse: Swets & Zeitlinger.

Vleuten, C.P.M. van der, Luyk, S.J. van & Beckers, A.J.M. (Ter perse) A written test as an alternative to performance testing. *Medical Education.*

Vleuten, C.P.M. van der, Robroek, W.C.L. & Luyk, S.J. van (In voorbereiding) Assessing practical nursing skills: Three methods compared.

Williams, R.G., Barrows, H.S., Vu, N.V., Verhulst, S.J., Colliver, J.A., Marcy, M. & Steward, D. (1987) Direct, Standardized Assessment of Clinical Competence. *Medical Education, 21*, 482-489.

Wijnen, W.H.F.W. (1971) *Onder of Boven de Maat.* Academisch proefschrift Rijksuniversiteit Groningen.

HOOFDSTUK 3

TRAINING AND EXPERIENCE OF EXAMINERS

Summary

Variation in the accuracy of examiner judgments is a source of measurement error in performance-based tests. In previous studies using physician subjects, examiner training yielded marginal or no improvement in the accuracy of examiner judgments. This study reports an experiment on accuracy of scoring in which provision of training and background of examiners are systematically varied. Experienced faculty, medical students and lay subjects were randomly assigned to either Training or No-Training groups. Using detailed behavioural checklists, they subsequently scored videotaped performance on two clinical cases, and accuracy of their judgments was appraised.

Results indicated that the need for and effectiveness of training varied across groups: it was least needed and least effective for the faculty group, more needed and effective for medical students, and most needed and effective for the lay group. The accuracy of the lay group after training approached the accuracy of untrained faculty. Trained medical students were as accurate as trained faculty. For faculty and medical students, training also influenced the nature of errors made by reducing the number of errors of commission.

It was concluded that training varies in effectiveness as a function of medical experience and that trained lay persons can be utilized as examiners in performance-based tests.

Introduction

In the last ten years, use of performance-based testing methods has become widespread in medical education. In these assessment methods, examinees are presented with standardized challenges in simulated clinical situations and asked to demonstrate their skills. Examiners record the behaviour of examinees, and test scores are based on these judgments. Consequently, a source of measurement error in performance-based tests is variation in the accuracy of the judgments of examiners. This source of variation directly affects the precision of scores.

The impact of examiner variation in clinical examinations has been recognized for some time. Wilson et al., (1969) concluded:

> *"The wide observer variation in assessing the clinical competence of a candidate must be openly recognized by all who are charged with the responsibility of determining the result of an examination."* (p. 39).

These authors suggested that pass/fail decisions on the basis of clinical examinations should not be made, because of the magnitude of examiner variation.

One prominent solution to the problem of examiner variation was the introduction of more objective clinical examinations. By standardizing instructions, material, and patients and by formalizing criteria into detailed, behaviourally-anchored checklists, the arbitrariness of the clinical examination was reduced. Early exponents of these more objective examinations have been Harden & Gleeson (1979), Stillman et al., (1976) and Newble et al., (1978), among others. Their pioneering work has found world-wide application (Hart et al. 1986; Hart & Harden, 1987).

A complementary approach to reducing examiner variation is examiner training; it is the focus of this study. The need for and the effectiveness of examiner training has been the subject of some debate (Wakefield, 1985). Two major studies have specifically investigated the effect of examiner training in assessment of clinical performance[1].

Ludbrook & Marshall (1971) compared 8 trained and 8 untrained examiners conducting a traditional clinical oral examination (viva voce); all were experienced surgical staff members. The 2.5-hour training used a videotaped examiner-candidate-patient encounter. The correlation of examiner and co-examiner ratings was only slightly higher for the trained group: 0.55 versus 0.45, and there was no evidence that variation in pass/fail decisions was less among the trained examiners. The investigators concluded that the experiment was a "dismal failure" (p. 154), and that they saw little future prospect for the traditional clinical examination.

---

[1]Many more studies are directed to other disturbing factors in the observational process (cf. Wakefield, 1985) and observer variation outside the medical educational field (cf. Kent & Foster, 1977).

The Ludbrook & Marshall study, however, was conducted before more objective clinical examinations were introduced. Examiner ratings were fairly unstructured and subjective, and this may have influenced the results.

The second major study was conducted in the context of the modern objective clinical examination. Newble et al. (1980) reported a study of examiner training under well controlled circumstances, in which examiners rated performance on detailed checklists. Eighteen experienced clinical examiners (nine medical and nine surgical) were randomly assigned to three training conditions, six examiners each. The first group received no training, the second group received limited training (half hour) and the third group received more extensive (2 hours) training. Before training, all groups rated the performance of five videotaped students. Two months after training, examiners re-scored the same videotapes. Unfortunately, the three groups performed differently on the pre-training video-tapes: the examiners in the no-training group were most accurate, followed by the limited training group, with the extensive training group worst. Results from the post-training sessions indicated that training did not improve accuracy. The authors concluded that some examiners were inherently accurate and others were not; the former do not need training, and the latter are not improved by it. Thus, they viewed accuracy as an examiner characteristic and felt that objectivity of examinations is improved through careful selection of examiners, not training.

These results raise serious questions about the effectiveness of training. The effort put into training sessions may not be compensated by a gain in accuracy.

These studies have implicitly assumed that experienced faculty were necessary to obtain accurate ratings of performance. More recently, in order to cope with the logistics of large scale testing, lay persons have commonly been used to record examinee performance (Stillman et al., 1986, 1987; Petrusa et al., 1987; Williams et al., 1987). The effectiveness of training may well vary as a function of the experience and background of examiners. The present study investigated this possibility experimentally by systematically covarying provision of training and experience of examiners.

Methods

Subjects

Three groups participated in the study. The Experienced Group included 22 medical faculty (7 surgeons, 15 family doctors) who had previously been examiners in performance-based tests. Forty-one fourth-year medical students (in a 6-year medical school programme) formed an Intermediate Group; the medical students had completed most of the "Skillslab" programme in the medical school (Bouhuijs et al., 1987) and were familiar with the test format and the clinical skills to be rated. Forty students from the law and economics schools constituted the Lay Group; they had no medical background. All students were volunteers and received a (small) financial compensation for participation.

Subjects in the three groups were randomly assigned to Training and No-Training conditions. In the Lay Group, one extra student was erroneously assigned to the Training condition, resulting in 21 trained students and 19 untrained students.

To check on the comparability of the groups resulting from random assignment, medical school records of the Experienced and Intermediate Groups were investigated. The mean "observer accuracy index" (Van der Vleuten & Van Luyk, 1987) of faculty in the Experienced Group, across 25 previous test administrations (1982-1987), did not differ between the Training and No-Training conditions. In addition, the scores of fourth-year students allocated to Training and No-Training conditions did not differ significantly on the most recent assessment of their clinical skills on the Maastricht Skills-Test (Van Luyk & Van der Vleuten, 1986).

Materials

Experimental materials consisted of four professionally recorded videotapes, each depicting an examinee working through a case in a performance-based testing situation. Two tapes, lasting fifteen minutes each, were made for each of two cases. The Abdomen case involved an abdominal examination on a patient; the student examinee did not interview the patient, but merely performed the physical examination and reported the findings. In the Chest Pain case, the examinee interviewed the patient, examined the chest, and reported the findings.

For each of the cases, checklists were developed. The Abdomen checklist included 50 items, the Chest Pain checklist had 60 items; 27 of these concerned interviewing skills.

Procedure

The Training condition consisted of a practice session with two videotapes, one for each case. For each tape, general instructions were provided, the tape was run and scored, and group feedback on accuracy was provided. Sessions continued until group members felt adequately prepared. For the Experienced Group, this required one hour; for both other groups 1.5 hours were required.

Subjects in the No-Training condition were given five minutes to familiarize themselves with the checklists. A brief explanation of the medical terms was also provided to the Lay Group. All subjects rated the remaining two videotapes in random order.

A group of 6 Skillslab faculty members developed consensus ratings of the videotapes to serve as a scoring key. Accuracy was expressed as percentage agreement with the key and was calculated for all subjects.

Results

Table 1 provides means and standard deviations for all groups. Figure 1 provides descriptive information in graphical form.

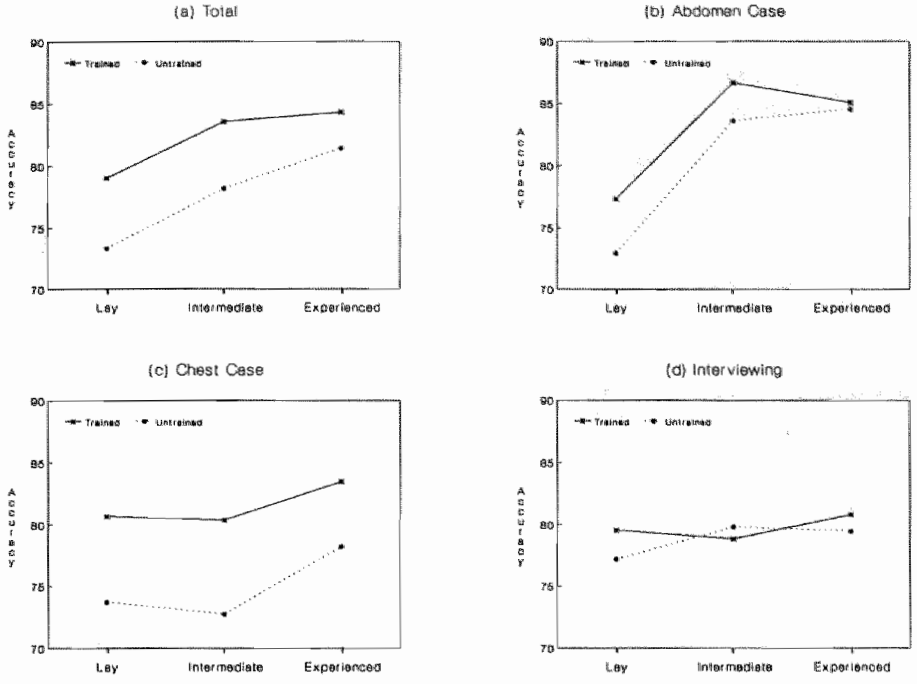

Figure 1: *Accuracy in relation to training and no-training conditions at three experience levels.*

Exploratory data analysis indicated that the scores on the interview portion of the Chest Pain checklist yielded a totally different pattern of results than the scores on the remaining items. Consequently, a separate subscale was constructed from these items for independent analysis.

Table 1: *Descriptive statistics for all groups*

	Sample Size	Abdomen Checklist		Chest Pain Checklist		Chest Pain Interview		Total[1]	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
No-Training	50	79.76	8.47	74.31	8.66	78.74	8.33	77.03	8.95
Experienced	11	84.55	4.66	78.24	8.11	79.46	8.34	81.39	7.21
Intermediate	20	83.60	5.75	72.73	7.68	78.62	9.87	78.16	8.67
Lay	19	72.95	8.31	73.69	9.64	77.19	6.56	73.32	8.88
Training	53	82.64	6.15	81.13	7.02	79.53	7.47	81.89	6.61
Experienced	11	85.09	4.93	83.47	5.31	80.81	7.91	82.28	5.07
Intermediate	21	86.67	3.37	80.38	7.59	78.84	7.87	83.52	6.61
Lay	21	77.33	4.99	80.67	7.26	79.54	7.08	79.00	6.38

[1]Excluding interview scores

Table 2: Results of analysis of variance of agreement percentages.

Source	SS	df	MS	F	p
Between subjects:					
Experience (E)	1471.53	2	735.77	13.32	0.00
Training (T)	1212.85	1	1212.85	21.96	0.00
E x T	111.11	2	55.55	1.01	0.37
Error	5357.92	97	55.24		
Within subjects:					
Cases (C)	603.60	1	603.60	16.10	0.00
E x C	1146.83	2	573.41	15.30	0.00
T x C	200.39	1	200.39	5.35	0.02
E x T x C	1.05	2	0.52	0.01	0.99
Error	3635.71	97	37.48		

A fixed-effects analysis of variance (ANOVA) was performed with Training condition and Experience level as between-subjects factors and Cases as within-subject factors (Winer, 1971, p. 563). Table 2 presents the results of the ANOVA.

The main effects for Experience, Training, and Case are all significant, indicating that training was effective overall, that experience also affected accuracy, and that cases differed in average level of accuracy. The Training-by-Case and Experience-by-Case interactions are also significant, indicating that there were differences between the two cases in relation to training and experience; this is well illustrated in Figure 1.

Figure 1A indicates that the performance of the Experienced Group under the Training condition is similar to the overall level of accuracy typical of the Maastricht Skills Tests (Van der Vleuten & Van Luyk, 1987). Figure 1A also indicates that the Intermediate Group is almost as accurate as the Experienced Group after training. The Lay Group, without training, is much less accurate than the two other groups. However, after training, Lay Group members perform better than untrained Intermediates and almost as well as untrained Experienced Group members.

Figures 1B and 1C depict results for cases separately. They indicate the source of the significant Experience-by-Case interaction in the ANOVA. Training in the Abdomen case yields only modest improvement for the Intermediate Group and no improvement for the Experienced Group. While the difference between these groups is small, the Lay Group performs much less accurately. For the Chest Pain case, Experience seems less important and Training has an overall effect. In other words, for the Abdomen case, experience matters, and, for the Chest Pain case, training is influential.

There are several possible explanations for the differences between the cases. Compared with the Chest Pain case, more speed was required to complete the Abdomen checklist, which may have handicapped Lay subjects. Alternatively, differences in the checklists for the cases may have produced the differences. The Abdomen checklist was more global than the Chest Pain checklist, making

interpretation more important in the Abdomen case. Conversely, less complex interpretation in the Chest Pain case could give rise to the dominant training effect.

Figure 1D depicts results for the interviewing portion of the Chest Pain case. Neither experience nor training had an influence, and the absolute accuracy reached (75%) was reasonably satisfactory. Since interviewing items required simple recording of the verbal information on the tape, medical experience apparently was not required.

Accuracy has, thus far, been defined by the simple percentage agreement of observer scores with the key. However, two kinds of errors are possible: errors of commission (giving credit when not justified) and errors of omission (not giving credit when justified). It was expected that subjects with less experience and/or training would see more adequate behaviour than actually present, whereas the experienced and/or trained subjects would be more critical. Table 3 presents the results of the error analysis.

Inspection of the table confirms this expectation, except for interviewing, where only errors of omission were made (all subjects marked less than indicated on the key). Lay subjects, whether trained or untrained, make more errors of commission. In the No-Training condition, more errors of commission are also produced by Intermediate and Experienced subjects. For the Training condition, this difference disappeared; Experienced subjects make more omission errors. In general, training seems to reduce errors of commission.

Table 3: Number of errors (O = omission; C = commission) in relation to training and experience (1 = Lay Group; 2 = Intermediate Group; 3 = Experienced Group).

Training Condition	Experience Level	Error Type	Total[1]	Abdomen Case	Chest-Pain Case	Chest-Pain Interview
	1	O	2.79	3.33	2.24	3.74
		C	5.38	7.10	3.67	1.52
Training	2	O	3.00	3.52	2.62	4.43
		C	3.33	3.14	3.52	0.76
	3	O	3.64	4.64	2.64	4.18
		C	2.77	2.18	2.72	1.00
	1	O	3.79	5.53	2.05	4.73
		C	6.49	7.42	5.37	1.53
No-training	2	O	1.63	2.50	0.75	3.25
		C	5.22	5.00	5.45	0.65
	3	O	2.64	4.00	1.27	3.82
		C	4.77	3.73	5.18	0.64

[1]Excluding interview scores

Discussion

This experiment showed that training has a statistically significant effect on the accuracy of observers scoring clinical examinations. Experience, however, moderates the magnitude of this effect.

Overall, the gain in accuracy was smallest for the most experienced group. This is in accordance with prior research showing modest (or no) gains through training. Scoring accuracy of experienced subjects was reasonably good, and only marginal gains in accuracy were achieved through training. Lay subjects gained most through training. Although their performance remained lowest, scoring accuracy approached the level of untrained experienced examiners. Depending upon the level of accuracy required, it appears that trained lay people can be used as examiners.

Scoring of interviewing appeared to be straightforward. Training and experience were not related to accuracy, and the absolute level of agreement was satisfactory. This was probably the effect of the use of content items: process-oriented items pertaining to the quality of the interview, which are also used in regular Maastricht performance tests (Kraan & Crijnen, 1987), may yield different results.

The analysis of error-type provided more qualitative information about the relationship between accuracy, training, and experience. Training reduces errors of commission and slightly increases errors of omission for subjects with some medical background. For lay subjects, it reduces both types of errors.

Some differences between cases were observed suggesting other moderating factors. First, speed of scoring may be important: scoring under time pressure may favor experience. Second, the nature of the checklist may play a role. The checklists used in the study were rather detailed and explicit, requiring little inference. Consequently, experience may have little influence, and training may be more effective. With checklists of a more global nature, the opposite may occur: subtle inferences may be required, and experience may matter more.

From a practical perspective, this study showed that lay people can be utilized as examiners in clinical examinations. Individuals without any medical background can reach a fair level of accuracy. The frequent use of standardized (simulated) patients for scoring performance-based tests (Williams et al. 1987; Petrusa et al. 1986; Stillman et al. 1986; Stillman et al. 1987) is supported by this outcome. This is particularly true for long multi-case tests, in which many observers contribute to the ratings (Swanson, 1987), since rating errors should "average out" across cases. Reproducibility of performance-based test scores is affected by both variation in examiners *and* variation across stations. Recent psychometric analyses have indicated that the latter is the more important source of measurement error and that many stations are required to achieve adequate reproducibility (Swanson, 1987; Van der Vleuten et al. 1988). Given the results of this study, training of more available and less expensive lay examiners may compensate for the limited loss of accuracy in scoring.

References

Bouhuijs, P.A.J., Vleuten, C.P.M. van der & Luyk, S.J. van (1987) The OSCE as a part of a systematic skills training approach. *Medical Teacher, 9*, 183-191.

Elstein, A., Shulman, L.S. & Sprafka, S.A. (1978) *Medical Problem Solving: An Analysis of Clinical Reasoning*. Harvard University Press, Cambridge Massachusetts.

Harden, R.M. & Gleeson, F.A. (1979) ASME Medical Education Booklet No. 8. Assessment of clinical competence using an objective structured clinical examination (OSCE), *Medical Education*.

Hart, I.R., Harden, R.M. & Walton, H.J. (Eds.) (1986) *Newer Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Hart, I.R. & Harden, R.M. (Eds.) (1987) *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.

Kent, R.N. & Foster, S.L. (1977) Direct observational procedures: Methodological issues in naturalistic settings. In: Ciminero, A.R., Calhoun, K.S. & Adams, H.E. (Eds.) *Handbook of Behavioral Assessment*. New York: John Wiley & Sons.

Kraan, H. & Crijnen, A. (1987) *The Maastricht History Taking and Advice Checklist: Studies of Instrumental Utility*. Unpublished doctoral dissertation.

Ludbrook, J. & Marshall, V.R. (1971) Examiner training for clinical examinations. *British Journal of Medical Education, 5*, 152-155.

Luyk, S.J. van & Vleuten, C.P.M. van der (1986) The assessment of clinical and technical skills at the medical school of Maastricht. In: Hart, I.R., Harden, R.M. & Walton, H.J. (Eds.), *Newer Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Newble, D.I., Elmslie, R.G. & Baxter, A. (1978) A problem based criterion-referenced examination of clinical competence. *Journal of Medical Education, 53*, 720-726.

Newble, D.I., Hoare, J. & Sheldrake, P.F. (1980) The selection and training for clinical examinations. *Medical Education, 14*, 345-349.

Norman, G.R. & Tugwell, P., Feightner,J.W., Muzzin, L.J., & Jacoby, L.L. (1985) Knowledge and clinical problem solving. *Medical Education, 19*, 344-356.

Petrusa, E.R., Blackwell, T.A., Rogers, L.P. & Saydjari, C., Parcel, S. & Guckian, J.C. (1987) An objective measure of clinical performance. *The American Journal of Medicine, 83*, 34-42.

Stillman, P., Sabers, D. & Redfield, D. (1976) The use of paraprofessionals to teach and evaluate interviewing skills in medical students. *Pediatrics, 57*, 769-774.

Stillman, P., Swanson, D., Smee, S., et al. (1986) Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine, 105*, 762-771.

Stillman, P., Regan, M., and Swanson, D. (1987) A diagnostic fourth year performance assessment. *Archives of Internal Medicine, 147*, 1981-1985.

Swanson, D.B. (1987) A measurement framework for performance based tests. In: Hart, I.R. & Harden, R.M., (Eds.), *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.

Vleuten, C.P.M. van der & Luyk, S.J. van (1987) Decomposition of OSCE's: Some methodological considerations and empirical findings. In: Hart, I.R. & Harden, R.M., (Eds.), *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.

Vleuten, C.P.M. van der, Luyk, S.J. van & Swanson, D.B. (1988) Reliability (generalizability) of the Maastricht skills test. *Proceedings of the Twenty-Seventh Annual Conference on Research in Medical Education (RIME)*, Chicago, USA.

Wakefield, J. (1985) Direct observation. In: Neufeld, V.R. & Norman, G.R. (Eds.) *Assessing Clinical Competence*. New York: Springer.

Williams, R.G., Barrows, H.S., Vu, N.V., Verhulst, S.J., Colliver, J.A., Marcy, M. & Steward, D. (1987) Direct, standardized assessment of clinical competence. *Medical Education, 21*, 482-489.

Wilson, G.M., Lever, R., Harden, R.M., Robertson, J.I.S. & MacRitchie, J. (1969) Examination of clinical examiners. *The Lancet, I*, 37-40.

Winer, B.J. (1971) *Statistical Principles in Experimental Design*. Tokyo: McGraw-Hill.

HOOFDSTUK 4

A WRITTEN TEST AS AN ALTERNATIVE TO PERFORMANCE TESTING

Summary

Performance tests are logistically complex and time consuming. To reach adequate reliability long tests are imperative. Additionally, they are very difficult to adapt to the individual learning paths of students, which is necessary in problem based learning. This study investigates a written alternative to performance based tests. A knowledge test of skills (KTS) was developed and administered to 380 subjects of various educational levels, including both freshmen and recently graduated physicians.

By comparing KTS-scores with scores on performance tests strong convergent validity was demonstrated. The KTS failed discriminant validity when compared with a general medical knowledge test. Also the identification of subtests discriminating between behavioural and cognitive aspects was not successful. This was due to the interdependence of the constructs measured. The KTS was able to demonstrate differences in ability level and showed subtle changes in response patterns over items, indicating construct validity. It was concluded that the KTS is a valid instrument for predicting performance scores and could very well be applied as supplementary information to performance testing. The relative ease of construction and efficiency makes the KTS a suitable substitute instrument for research purposes.

The study also showed that in higher ability levels the concepts which were meant to be measured were highly related, giving evidence to the general factor theory of competence. However, it appeared that this general factor was originally non-existent in freshmen and that these competencies integrate as the educational process develops.

Introduction

In recent years medical assessment literature has focussed on the assessment of clinical skills in semi-naturalistic settings (e.g. Hart, Harden & Walton, 1986). This interest probably originated from an overall dissatisfaction with the traditional written and oral assessment formats, which are supposed to be unable to measure more complex professional skills. Tests simulating real life situations for students are considered to have more fidelity. The Objective Structured Clinical Examination (OSCE) developed by Harden & Gleeson in 1979, and tests with simulated patients (Stillman, Sabers & Redfield, 1976), among others, initiated a trend for many educationally involved faculty to introduce comparable instruments in their curriculum.

The empirical results on the validity of these performance tests support their use. However, the major drawback seems to be the content specificity of the skills assessed (Swanson, 1988). The content of the tasks or cases presented to students appears to induce high variability in the scores obtained (e.g. Elstein et al., 1978; Norman & Tugwell., 1982). As a result, to reach generalizable scores over occasions students have to complete many cases. Consequently, testing time may take up to 4 hours (Stillman, et al., 1987) or as much as 15 hours (Williams et al., 1987).

The Maastricht medical school uses a similar naturalistic testing format to test the proficiency of students in technical and clinical skills. The achievements of these skills are considered very important in the educational programme. The integration of theory and practice is heavily emphasized within the context of the Maastricht problem based learning curriculum (Schmidt, 1983). By means of a laboratory situation students learn technical and clinical skills directly from the beginning of their studies, integrated in the curriculum. In the first 4 years, students spend approximately 2-3 hours per week in this Skillslab. They are guided by specially trained Skillslab personnel (Physicians, physiotherapists, laboratory specialist, social scientists etc.). Year 5 and 6 consist of clinical rotations, at which time the students are well prepared and able to apply their skills to everyday practice.

At the end of each academic year the proficiency of all students (including classes 5 and 6) are assessed with a Skills Test. In this OSCE-like instrument students are tested for 2 hours on a number of stations (6-12), varying from 10 to 20 minutes each. Trained medical faculty members score the behaviour of students on detailed checklists. About 900 students are tested each year in this way. The test is completely restricted to 'hands on' performance and unlike most OSCE's it contains no written parts (cf. Van Luyk et al., 1986; Bouhuijs et al., 1987).

Logistically this semi-naturalistic assessment procedure is costly. Although financial burdens are not too abundant (e.g. Williams et al., 1987), a smooth organization of performance tests is complex and consumes a lot of faculty time. Total testing time and frequency are therefore limited. A Maastricht student is only tested once a year for 2 hours, which can really be seen as a minimum. Limited time is additionally troublesome, since content specificity, as was delineated above, is a major problem and requires long tests. Therefore

valid alternative instruments with less demanding characteristics would be very useful. Additional to the information gathered from performance tests, alternative instruments could be used for supplementary information yielding overall more reliable results.

A second reason for seeking alternative measures to performance testing concerns an educational issue, related to the restricted sampling characteristics of performance tests and the specific demands of the Maastricht medical curriculum. In problem-based learning the principle of self directed learning (Barrows, 1980) is fundamental. It implies that students themselves are responsible for their own study actions (as is expected in their later professional life) and that within the boundaries provided by the thematically partitioned curriculum, students are free to choose study topics. In practice this means that different students study different material at the same educational level within the same course. To reward these individual (or group) learning paths the evaluation system should be adapted accordingly. This can be achieved by means of 'Progress Testing' (Verwijnen et al., 1982). In this technique the entire student body of a school is periodically (e.g. four times a year) submitted to a large test of (multiple) true/false questions covering the entire field of medicine, together reflecting the end objectives of a curriculum.

Progress Tests are adapted to self-directed learning. Individual learning activities are reinforced (samplewise) since the test assesses the entire field of medicine. Conversely, a student cannot be tempted to study specifically for the test. Specific preparation is useless since the exact content of the next test is unknown. In progress testing the direct relationship between direct prior education and the content of a test is cut. The technique also possesses a number of additional educational advantages which will not be outlined here (cf. Verwijnen et al., 1982).

However, this wide sampling over an entire competence domain, essential for Progress Testing, is impossible with performance tests. In these tests very efficient sampling is indicated. Stations or cases are the smallest testing units comparable to multiple choice items in a written test, but they are much more time consuming. Performance tests normally have a limited number of stations due to the mentioned logistical limits. The selection of stations which content is not related to prior education (and which a student could 'skip' as in a Progress Test) is impossible. As a consequence students are fairly able to predict the content of a particular Skills Test and are able to specifically prepare themselves (or even cram) for the test: the test becomes the steering factor, not the student himself. This extrinsic motivational factor should be avoided especially in problem based learning. An alternative measure in which these sampling problems do not occur, would thus benefit the educational goal of self-directed learning.

Knowledge test of skills

The study reported here will investigate such an alternative instrument. It concerns an experimental administration of a written test focussing on the knowledge of technical and clinical skills. The test contains simple objective questions (true/false items) constructed as a Progress Test. Questions in the test may focus on the cognitive part of technical and clinical skills, but may also

simulate 'hands on' items of a performance test checklist. The first type of questions refer more to the thinking or decision making process, whereas the second type of questions pertain more to the procedures of skills. The following is an example of an item directed to the thinking process: 'A negative response to the Rinne test is an indication of a sensorineural hearing loss'. A more behaviourally oriented item is: 'When bandaging the knee the first sleeve is applied above the knee'.

The choice for a knowledge test was based on its feasibility. Knowledge tests are relatively easy to construct and broad and easy sampling over content areas is possible. In addition, the choice was motivated by the theoretical intrinsic relationship between knowledge of skills and the performance of these skills (Gagné, 1977; Ebel, 1975 (in Maatsch & Huang, 1986), Glaser, 1984 and 1987). The exact nature of this relationship, e.g. the organization of knowledge (Norman et al., 1985), is still a matter of scientific study, but there is no doubt that knowledge plays an important role for adequate performance in technical and clinical skills. In discussing the variability problem of performance measures Norman et al. (1985) concludes:

> *'Alternatively, one may have to accept that this degree of variability is to be expected, and devise new approaches to assessment which more directly top into the extensive body of knowledge which is the hallmark of clinical expertise.'* (p. 354).

The present study focusses on such an alternative approach. The value of the *Knowledge Test Of Skills* (KTS) will be investigated by assessing its reliability and by relating it to a number of validity criteria.

Validity questions

Four research questions were posed with regard to the validity of the KTS:

1. Can the KTS predict results on performance tests (convergent validity)?
2. Is the KTS capable of discriminating from a non-intended competency (discriminant validity)?
3. Are sub-tests identifiable referring more to the cognitive domain or to the behavioural domain (construct validity)?
4. Can it discriminate between groups of different levels of ability (construct validity)?

The first question will be addressed by correlating the KTS with the Skills Test.

Discriminant validity will be assessed by comparing KTS-results with the regular Maastricht Progress Test (MPT). Four times a year this general medical knowledge test is completed by all students of the medical school. Each test, parallel in content, is newly constructed (or retrieved from an itembank). The scores on the MPT are very appropriate as a criterion measure for discriminant validity. The KTS is supposed to measure a specific trait also assessed by performance tests such as the Skills Test, and not, or to a lesser extent, to assess general medical knowledge. Therefore low correlations of KTS and MPT scores should appear.

To address the third question, test items will be rated by experts as referring to either cognitive or to behavioural aspects, thus forming extreme scales

within the KTS. Subsequently, with the help of a statistical technique maximum unidimensionality of these scales will be achieved.

The final question will be dealt with by comparing scores over different ability groups, including two reference groups of freshman and physicians, on the basis of their composite scores on the KTS, as well as on their item score pattern.

Methods

Instruments

The KTS consisted of 238 questions in the (multiple) true/false format with a question mark option. Together these items constituted a sample of the complete instruction programme of the Skillslab. They were balanced according to a blueprint of 10 skill-domains, similar to the blueprint of the Skills Test. It was constructed by the faculty of the Skillslab and reviewed by a committee of three physicians.

The Skills Tests (SKTs) used in this study consisted of the regular SKTs of the evaluation programme. They consisted of a varying number of stations (between 7 and 11) attuned to the educational level (class). Stations were sampled on the basis of a fixed blueprint, and testing time was fixed to 2 hours per student.

The Maastricht Progress Tests used here consisted of 250 to 300 questions, also in the true/false format and question mark option. It is structured according to a fixed blueprint of mainly organ systems.

Subjects

From each of the six classes volunteers were asked to complete the KTS. In total 244 students responded (n subjects of class 1 to 6 with percentage of total year group in brackets: 40(25%), 83(60%), 44(40%), 26(26%), 24(38%), 27(47%)). They received a small financial compensation for their efforts. Two reference groups were used. First, a group of 61 freshmen were taken to represent a group of novices. They completed the KTS in the second week following their entrance in the medical school. The total number of these 305 students represented 47% of the total student population at that time. The second reference group consisted of a national sample of 75 recently graduated physicians following their specialty training in family medicine. This group was considered to provide an external reference representing the end level of proficiency to be reached. In total 380 subjects completed the KTS.

Since SKT and MPT are part of the regular evaluation system, from all students except the freshman group SKT and MPT results were available (also from the non-response group). Naturally, no SKT and MPT scores from the physician group were available.

Procedure

The administration of the KTS took place in 1984, separate for each participating group. The freshmen group was tested in September, the second week after entrance. For the classes 1 to 6 the administration took place within one week after completion of their regular Skills Test, all in the period between March and June 1984. For each class the Skills Test was administered at a different point in time, all somewhere at the end of the academic year. The dispersed KTS administration was preferred above a single testing moment, to achieve a constant time-lag between KTS and Skills Test completion for all classes. To avoid copying all test-booklets were collected after completion. In class 5 and 6 the original response appeared to be low. The test-booklets were distributed to students at home with the specific instruction not to consult any text books. The resulting response was satisfying, but standard conditions are not guaranteed. All other administrations took place under regular standard test taking conditions. The reference group of physicians completed the test in the context of their educational programme in the period from November to December 1984.

Since the KTS administration was preceded by the Skills Test for that specific class, time between KTS and Skills Test was constant. Time constancy was not met for Progress Test data, since the MPT is administered to all students at the same time. Therefore two MPT's were taken (March and June) and student scores were averaged. This procedure was also warranted since correlations between successive Progress Tests are high (in this case the median correlation over the six classes was 0.84).

Unlike SKT and MPT the KTS had no consequences for the students in terms of decision making. For all tests the percentage correct was calculated and used.

Results

Sample representativeness

A possible selection effect may have occurred by using volunteers. To check whether the sample of students participating in the KTS experiment was representative of the total student body, the achievements on the SKT and MPT of the response group, the non-response group and the total class were compared. Table 1 contains these results. Both reference groups are left out, since no SKT and MPT results were available from these participants.

Although slight differences in the usual direction of higher scores for the response group exist some of which are significant, table 1 shows that there are no large discrepancies. The sample of students who completed the KTS therefore seems adequately representative.

Table 1: *Means, standard deviations and range in percentage correct on the Skills Test and Maastricht Progress Test for the response group (R), non-response group (N) and total group (T).*

			Skills Test			Maastricht Progress test		
Class	Group	n	m	s	min-max	m	s	min-max
1	R	40	81	6	65-90	16	5	9-31
	N	117	77*	7	57-90	14*	5	7-31
	T	157	78**	7	57-90	15	5	7-31
2	R	83	70	6	55-87	28*	7	14-53
	N	55	69	8	47-85	31	9	14-57
	T	138	70	7	47-87	29	8	14-57
3	R	44	67	8	51-81	36	7	24-54
	N	66	66	8	47-82	35	8	19-56
	T	110	66	8	47-82	35	7	19-56
4	R	26	73	7	56-84	46	10	30-68
	N	73	72	10	42-88	44	8	21-62
	T	99	72	9	42-88	45	9	21-68
5	R	24	70	8	44-81	55	7	39-68
	N	39	67	9	36-83	51*	7	29-65
	T	63	68	9	36-83	52	7	29-68
6	R	27	70	7	56-83	61	9	42-75
	N	31	68	8	50-81	55*	7	36-69
	T	58	69	8	50-83	58	8	36-75

* significant t-value ($p < 0.05$) between response and non-response group
** significant t-value ($p < 0.05$) between response and total group

Reliability

For each group the reliability of the KTS was estimated by calculating a generalizability coefficient (Brennan, 1983). The results are given in table 2.

Table 2: *Generalizability coefficients of knowledge-test-of-skills per participating group (0 = freshmen; 1-6 = year 1 to 6; 7 = physicians).*

	Group							
	0	1	2	3	4	5	6	7
Generalizability	0.95	0.93	0.90	0.90	0.90	0.92	0.92	0.89

The coefficients all exceed the standard benchmark of 0.80. It can be concluded therefore that the KTS supplies reliable information, corresponding to regular MPT results and other multiple choice or true/false tests.

Validity

To address the first question of convergent validity the first column of table 3 contains the correlation per class between the KTS with the Skills Test.

Table 3: Observed and disattenuated[1] correlations (left and right value resp.) between instruments used.

	Correlation between		
Class	KTS-SKT	KTS-MPT	SKT-MPT
1	0.03/0.04	0.27/0.30	0.19/0.24
2	0.27/0.33	0.51/0.56	-0.05/-0.06
3	0.60/0.74	0.37/0.42	0.10/0.12
4	0.45/0.55	0.37/0.41	0.68/0.84
5	0.51/0.65	0.41/0.48	0.66/0.82
6	0.72/0.89	0.60/0.66	0.60/0.74

[1]Reliability used for SKT was the overall found generalizability value of the SKT of 0.73 (cf. Van der Vleuten & Van Luyk, 1988); for MPT generalizabilities per class were estimated after pooling variance components of the two MPT's used.

The correlations with SKT generally shows an incline over years, starting with a low value for the first year to a very high value in the last year. The correlation corrected for unreliability in year 6 is nearly perfect. Apparently a score on the written KTS can well predict the performance score on the SKT, especially in higher year groups. The low correlations in the lower classes cannot be attributed to a simple bottom effect, since the variances of test scores in all classes appeared to be about equal.

Discriminant validity would be demonstrated if the KTS would measure a different trait from the general medical MPT. From the second column of table 3 this seems not to be the case. Although the correlations are somewhat lower then the corresponding ones of column 1 (except for year 1 and 2), the correlations are far from zero. In addition, with the exception of year 2 the same pattern of incline in correlation appears. With these moderate to high correlations it must be concluded that the KTS also predicts the ability in general medical knowledge. It either implies that the KTS assesses multiple traits or that the supposed different traits are correlated (trait variance). From the third column of table 3 the latter explanation seems more likely. The SKT and MPT, although intended to measure very different competencies, also show considerable correlations.

The third validity question elaborates on this relationship. Despite the relatively high intercorrelations there is still a possibility that part of the KTS is responsible for the correlation with general medical knowledge and another distinct part accountable for correlations with the behavioural part. If these subtests could be identified, it would be a strong indication for construct validity of the KTS, supporting the cognitive/behavioural dimension in technical and clinical skills.

To identify these sub-tests two steps were taken: a content analysis and a statistical procedure. After receiving explicit instructions and training, seven raters (five physicians and two social scientists, familiar with the Skillslab programme) were asked to rate each item as behavioural (category 1 items), as cognitive (category 2 items) or indifferent. Consensus appeared to be low (overall kappa 0.40; Cohen, 1960). By removing two raters (cf. Schouten, 1982) the agreement increased to 0.54. If four of these five raters agreed, an item was placed in one of the three categories. This resulted in 42 items in category 1 and 51 in category 2. Subsequently in the second step these items were analyzed to fit in a unidimensional scale by means of Rasch analysis (Rasch, 1960). In this analysis responses to items are tested against a theoretical response model or item characteristic curve. Item response theory is meant as an alternative to classical test theory. The Rasch model is one of the strongest item response models, but also the severest one. Items forming a scale are supposed to have identical item-characteristic curves, only differing from each other in difficulty (cf. Hambleton & Cook, 1977). The technique was applied here because items forming a Rasch homogeneous scale are by definition unidimensional (otherwise a fit will not be found). The rationale was that items considered as belonging to one extreme of a dimension on the basis of their contents would maximally be unidimensional if held against the demands of this statistical technique. In this way it was attempted to diverge the two extreme scales hypothesized in the KTS (the category 1 and 2 items). For category 1 this resulted in 12 items (Anderson test: $X^2 = 21.00$; df = 11, p = 0.033), in category 2 18 items remained ($X^2 = 14.72$; df = 17, p = 0.648).

Total scores on these new scales were correlated with SKT and MPT. Category 1 scores should correlate highly with SKT scores and low with MPT, whereas the reverse should hold for category 2. In table 4 these relations are shown.

Table 4: Correlations of category 1 (behavioural items) and category 2 (cognitive items) with Skills Test (SKT) and Maastricht-progress-test (MPT).

	Category 1 with		Category 2 with	
Class	SKT	MPT	SKT	MPT
1	0.10	0.15	-0.11	0.24
2	0.26	0.25	0.02	0.28
3	0.54	0.17	0.55	0.23
4	0.07	0.18	0.16	0.15
5	0.31	0.50	0.10	0.03
6	0.20	0.34	0.44	0.26

Most correlations are low. The few higher correlations scatter in all directions. There is no evidence for either scale, indicating that these sub-tests did not succeed in differentiating between a more cognitive and a more behavioural dimension in the KTS.

The fourth validity question of KTS-differences between varying levels of ability, can in the first place be demonstrated by comparing mean performance of the different participating groups. Figure 1 depicts these results. The standard deviation of scores is drawn above and below each average.

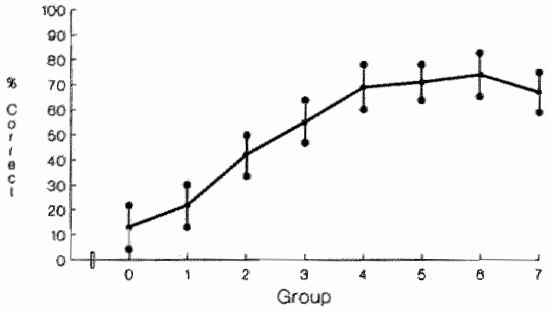

Figure 1: Mean scores and standard deviations on the KTS of different groups of ability level (0 = novices, 1-6 = classes 1-6, 7 = family medicine residents).

Although the means in the figure are connected, it should be realized that these data are cross-sectional, albeit regular Progress Test longitudinal growth curves are more or less comparable (cf. Verwijnen et al., 1982). There is a clear incline in proficiency in the first 4 years (statistically significant on the basis of non-overlapping confidence intervals ($p < 0.05$) of the means), leveling off in year 5 and 6. The reference group of physicians score on the same level as class 4. The stabilizing competency in year 5 and 6 is different from regular Progress Tests assessing general medical knowledge. Perhaps the KTS is insensitive at this proficiency level, or, a more tentative conclusion, the instructional programme of the Skillslab reached the intended goal of fully preparing students for their clerkships.

Mean performance however may conceal different response patterns. Investigating changes at the item level would be a more subtle approach to differentiate competency levels. Taking the reference group of family medicine residents as 'gold standard' the hypothesis can be made that with increasing competency the score pattern on items of higher classes should match the pattern of the reference group.

The p-values (proportion of persons that answered an item correctly) within every group were classified into four equal parts: Category 1: $0.00 \leq p < 0.25$; category 2: $0.25 \leq p < 0.50$; category 3: $0.51 \leq p < 0.75$ and category 4: $0.75 \leq p < 1.00$. Category 1 thus represents the very difficult items, category 4 the very easy items answered correctly by most people. Subsequently for each class cross tabulations were made of frequencies of items falling in these categories and related to the frequencies of the reference group. This led to the six contingency tables given in table 5.

Each contingency table is horizontally percented to make chance inferences. For example, for class 1 it can be said that the chance of an item being easy for first-year students and difficult for physicians is 0% (which makes sense); conversely an item being difficult for first year students has a chance of 9% of being difficult for the reference group. If the reference group is taken as a gold standard then the patterns over the successive years should converge, leading to empty off-diagonal elements in higher years and fuller diagonal frequencies.

Table 5: *Contingency tables of item difficulties of every class compared with the reference group of physicians (1 = 0.00 < p < 0.25; 2 = 0.25 < p < 0.50; 3 = 0.51 < p < 0.75; 4 = 0.75 < p < 1.00).*

		Physicians				
		1	2	3	4	Σ
Class 1	1	9	21	34	36	100
	2	8	16	13	63	100
	3	6	11	39	44	100
	4	0	6	24	71	100

		Physicians				
		1	2	3	4	Σ
Class 2	1	14	24	38	24	100
	2	7	22	27	45	100
	3	2	17	34	46	100
	4	2	2	16	80	100

		Physicians				
		1	2	3	4	Σ
Class 3	1	26	24	32	18	100
	2	11	32	38	20	100
	3	3	17	36	44	100
	4	0	3	16	81	100

		Physicians				
		1	2	3	4	Σ
Class 4	1	43	29	19	10	100
	2	14	41	33	12	100
	3	3	25	40	32	100
	4	2	4	21	68	100

		Physicians				
		1	2	3	4	Σ
Class 5	1	61	39	0	0	100
	2	19	45	36	0	100
	3	3	25	46	25	100
	4	0	4	25	71	100

		Physicians				
		1	2	3	4	Σ
Class 6	1	69	31	0	0	100
	2	30	59	11	0	100
	3	3	25	54	18	100
	4	0	3	22	76	100

Close inspection of table 5 shows that this is indeed the case. The response pattern of class 6 is almost identical to the reference group. In addition, unlike the mean increase from figure 1, this proximity is not maximally reached in class 4 but continues in year 5 and 6. This means that growth of competency in the clinical period of year 5 and 6 is not reflected in mean level of proficiency, but in a differently composed proficiency. With increasing competence different parts of the KTS questions are answered correctly, reflecting gradually the answering pattern of physicians. Students in the clinical period probably forget (irrelevant) material (judged by the reference group) and learn other (clinically relevant) material. This can be interpreted as an indication for validity of the KTS.

Discussion

The reliability of the KTS was found to be more than adequate. Naturally this finding is not very surprising since high reliabilities are common with true/false (and multiple choice) tests (e.g. Norcini et al., 1985).

With regard to validity the outcomes are more complex. Convergent validity appeared to be (very) high in higher classes. The (true) correlation found of 0.89 in year 6 is even nearly perfect, indicating that the constructs measured with both tests are highly correlated and that written tests of knowledge are potentially able to predict scores on a performance test. At the high proficiency region the written test and the performance test are nearly interchangeable. For these so very different formats this is a very striking result.

Interestingly herein is the steady incline in correlation with increase of class, as if the concepts of knowledge and performance gradually integrate with growing proficiency.

On the other hand, discriminant validity could not be demonstrated. Although correlations of the KTS with a general medical knowledge test were somewhat lower, the same pattern of correlations appeared: low correlations in lower classes, high(er) correlations in the upper classes. The fact that the KTS correlated with both criterion measures could not be explained by the possibility that different parts were responsible for these relationships. The search for sub-tests related to behavioural and cognitive aspects did not succeed. Raters perceived difficulty in categorizing the items in these dimensions and their agreement appeared low to moderate. Also the high correlation of the Skills Test with the general medical knowledge test may account for the failure of identifiable sub-tests. It appeared that the traits assessed are far from independent, and rather highly correlated.

The KTS was able to discriminate between groups of different ability. A steady 'growth' in competency was seen until the fourth class where mean growth leveled off. However, analysis at the item level revealed shifts in response pattern in year 5 and 6 resembling more and more the answering pattern of the reference group of physicians. The KTS seems sensitive enough to assess these more or less subtle changes, indicative of the 'ability difference' aspect of construct validity.

The progress-testing ability of the KTS is demonstrated by the ability of the KTS to discriminate between different educational levels.

In summary, the written test on knowledge of skills is able to predict achievements in performance tests, except for students in low proficiency regions such as in the first and second year. This ability appears not to be unique for this specific instrument since a general medical knowledge test also correlated substantially. The tentative conclusion to be drawn for construct validity, is that the test seems valid but the constructs are highly related.

This latter result is also found more often in current literature. A number of studies have shown high correlations between very different formats intended to measure different constructs. Maatsch (1980, 1986, 1987) showed that with different instruments such as multiple choice tests, pictorial MCQs, PMPs, simulated patients, oral examinations and computer simulations the same, or very highly related, latent variable was assessed. High correlations with multiple choice has also been found in other studies, e.g. with regard to PMPs (Norcini et al.,

1985, 1986b), computer simulations (Norcini et al., 1986a) and open ended item formats (Norman et al., 1987). These empirical findings gave rise to the 'lump theory', stating that observed achievements on different tests are caused by one and the same competency, comparable to the g-factor in intelligence.

The results of this study fit in part into this theory. For the upper region of competency the traits are probably highly interrelated, forming a 'lump'. Most of the mentioned studies compare item formats in this upper ability region (e.g. board examinations, final examinations) and hardly ever, as was done in this study, in the beginning of a curriculum with students in the lower ability region. However, in lower classes the traits seem (nearly) independent. It suggests that the supposed g-factor is one which evolutes: different competencies are originally independent and integrate to a common denominator when proficiency or education progresses.

On the other hand it should be stressed that high correlations do not necessarily imply causal relationships and that instruments may rank order students in the same way but still measure different competencies. It should also be realized that even with a perfect correlation different instruments may still be educationally useful (Frederiksen, 1984). For instance, on the basis of the above data one could suggest replacing the (relatively expensive) Skills Test in higher classes by the (relatively cheap) knowledge test of skills, since the same psychometric information is gathered. However the correlations found are what they are because of the *existence* of a Skills Test. One of the most important characteristics of a testing instrument is its impact on the student's learning (as important as reliability and validity). A student will adapt his learning style to the testing method he will be submitted to. The replacement of a performance test by a knowledge test of skills, would elicit another study approach to the learning of technical and clinical skills. This approach is probably more cognitively oriented and students will likely focus on the mere knowledge in stead of the practice of actually doing. Without doubt, this is not the intention of present day educators.

In all, the KTS seems a valid instrument to assess technical and clinical skills and may well be used in assessment situations. Applying this kind of written testing for these kind of skills, may help somewhat to overcome the most eminent problem of performance tests. As stated earlier, to reach adequate reliability of performance tests, many cases have to be used leading to very long testing time and high cost. A KTS-like instrument supplying valid additional information may supplement performance tests, thus giving more reliable composite information against relatively lower cost and logistics. In addition, this kind of written testing can successfully be applied in a progress-testing format, which is for example important in problem based learning curricula such as the Maastricht medical school.

Another meaningful application of 'written performance testing' may lie in scientific educational research, e.g. for comparative purposes (e.g. Petrusa, 1987). For research goals the logistics of performance testing is in most cases a major obstacle, especially in larger scale projects. The more economic alternative paper and pencil test, as developed here, may still supply valid information.

References

Barrows, H.S. & Tamblyn, R.M. (1980) *Problem-based learning.* New York: Springer.

Brennan, R.L. (1983) *Elements of Generalizability Theory.* Iowa: American College Testing Program.

Bouhuijs, P.A.J., Vleuten, C.P.M. van der & Luyk, S.J. van (1987) The OSCE as a part of a systematic skillstraining approach. *Medical Teacher, 9,* 183-191.

Cohen, J. (1960) A coefficient of agreement for nominal scales, *Educational and Psychological Measurement, 10,* 37-46.

Elstein, A., Shulman, L.S. & Sprafka, S.A. (1978) *Medical Problem Solving: An Analysis of Clinical Reasoning.* Harvard University Press, Cambridge Massachusetts.

Frederiksen, N. (1984) The real test bias: influences of testing on teaching and learning. *American Psychologist, 39,* 193-202.

Gagné, R.M. (1977) *The Conditions of Learning.* New York: Crofts & Co.

Glaser, R. (1987) Learning theory and theories of knowledge. In: De Corte, E., Lodewijks, H., Parmentier, R. & Span, P. (Eds.), *Learning and Instruction. European research in an international context.* (Studia Pedagogica) (Vol.1, p. 397-414). Leuven/Oxford: Leuven University Press/Pergamon Press.

Glaser, R. (1984) Education and thinking: the role of knowledge. *American Psychologist, 39,* 93-103.

Hambleton, R.K. & Cook, L.L. (1977) Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14,* 75-96.

Harden, R.M. & Gleeson, F.A. (1979) ASME Medical Education Booklet No. 8. Assessment of clinical competence using an objective structured clinical examination (OSCE), *Medical Education.*

Hart, I.R., Harden, R.M. & Walton, H.J. (1986) *Newer Developments in Assessing Clinical Competence,* Montreal, Heal Publications.

Luyk, S.J. van, Vleuten, C.P.M. van der & Peet, D.G.M. (1986) The assessment of clinical and technical skills at the medical school of Maastricht. In: I.R. Hart, Harden, R.M. & Walton, H.J. (Eds.) *Newer Developments in Assessing Clinical Competence,* Montreal, Heal Publications.

Maatsch, J.L. (1980) *Model for a criterion-referenced medical specialty test.* Final Report Grant No. HS-02038-02, Office of medical Education Research and Development Michigan State University.

Maatsch, J.L. & Huang, R.H. (1986) An evaluation of the construct validity of four alternative theories of clinical competence. *Proceedings of the Twenty-fifth Annual Conference on Research in Medical Education, Washington, DC.*

Maatsch, J.L. (1987) Theories of clinical competence: The construct validity of objective tests and performance assessments. *Paper presented at the International Conference on Evaluation in Medical Education, Beer Sheva, Israel.*

Norcini, J.J., Swanson, D.B., Grosso, L.J., Shea, J.A. & Webster, G.D. (1985) Reliability, validity and efficiency of multiple choice question and patient management problem item formats in the assessment of physician competence. *Medical Education, 19*, 238-247.

Norcini, J.J., Meskauskas, J.A., Langdon, L.O. & Webster, G.D. (1986) An evaluation of a computer simulation in the assessment of physician competence. *Evaluation in the Health Professions, 9*, 286-304.

Norcini, J.J., Swanson, D.B., Grosso, L.J. & Webster, G.D. (1986) The psychometric characteristics of some common item formats. In: I.R. Hart, Harden, R.M. & Walton, H.J. (Eds.) *Newer Developments in Assessing Clinical Competence*, Montreal, Heal Publications.

Norman, G.R., Smith, E.K.M., Powles, A.C.P., Rooney, P.J., Henry, N.L. & Dodd, P.E. (1987) Factors underlying performance on written tests of knowledge. *Medical Education, 21*, 297-304.

Norman, G.R. & Tugwell, P. (1982) A comparison of resident performance on real and simulated patients. *Journal of Medical Education, 57*, 708-715.

Norman, G.R. & Tugwell, P., Feightner, J.W., Muzzin, L.J., & Jacoby, L.L. (1985) Knowledge and clinical problem solving. *Medical Education, 19*, 344-356.

Petrusa E.R. (1987) Improving clinical performance assessment: A multi-institutional trail of the OSCE. *Proceedings of the Twenty-sixth Annual Conference on Research in Medical Education, Washington, DC.*

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Schmidt, H.G. (1983) Problem based learning: Rationale and description. *Medical Education, 17*, 11-16.

Schouten, H.J.A. (1982) Measuring pairwise agreement interobserver agreement when all subjects are judged by the same observers. *Statistica Neerlandica, 36*, 45-61.

Stillman, P., Sabers, D. & Redfield, D. (1976) The use of paraprofessionals to teach and evaluate interviewing skills in medical students. *Pediatrics, 57*, 769-774.

Stillman, P.L., Regan, M.B. & Swanson, D.B. (in preparation) *A diagnostic fourth year performance assessment.*

Swanson, D.B. (1988) A measurement framework for performance based tests. In: I.R. Hart & R.M. Harden (Eds.) *Proceedings of the Second International Conference on Newer Developments in Assessing Clinical Competence*, Ottawa, Canada.

Verwijnen, G.M. Imbos, Tj., Snellen, H., Stalenhoef, B., Pollemans, M., Luyk, S., Sprooten, M., Leeuwen, Y. van & Vleuten, C.P.M. van der (1982) The evaluation system of the medical school of Maastricht. *Assessment and Evaluation in Higher Education, 3*, 225-244.

Vleuten, C.P.M. van der & Luyk, S.J. van (1988) Decomposition of OSCE's: Some methodological considerations and empirical findings. *Proceedings of the Second International Conference on Newer Developments in Assessing Clinical Competence*, Ottawa, Canada.

Williams, R.G., Barrows, H.S., Vu, N.V., Verhulst, S.J., Colliver, J.A., Marcy, M. & Steward, D. (1987) Direct, Standardized Assessment of Clinical Competence. *Medical Education, 21*, 482-489.

HOOFDSTUK 5

A VALIDITY STUDY OF A TEST FOR CLINICAL AND TECHNICAL MEDICAL SKILLS[1]

Summary

The procedure of testing technical and clinical skills at the Maastricht School of Medicine in the Netherlands has an OSCE-like format. All students are once a year submitted to this procedure. To enhance objectivity, scoring in each station is restricted to detailed checklists of criterion behavior.

To establish concurrent validity of this so-called Skills Test a study was undertaken in which general impressions registrated by (expert) observers on a rating scale (global judgement) were compared with scores on the checklists of the Skills Test (analytical judgement). Next to a general impression, observer ratings were gathered of specific components of competent student behavior to explain possible discrepancies. These components were aspects of problem solving, attitudes, personality factors and performance technique. By comparing ratings of specific competencies with Skills Test outcomes more of the construct validity of the latter might be established.

Results showed no gross differences between the global and analytic judgment methods, in terms of highly and poorly rated performance. Somewhat surprisingly, however, discrepancies were observed between checklist-based scores and ratings of components in areas in which the Skills should be most effective (i.e. technique).

Since global ratings were gathered during regular Skills Test administrations by regular examiners after they had scored with checklists, completion of global ratings may have been biased. From control experiments this bias appeared negligible.

[1]This text is a slightly edited version of the original published version.

Introduction

Technical and clinical skills are important elements of the curriculum of the Maastricht School of Medicine, the Netherlands. From the very beginning this problem-based curriculum (Schmidt, 1983) trains students in practical and clinical skills. A special department, the so-called Skillslab, was set up in which students spend at least two or three hours a week. They acquire all relevant skills in pace and accordance with the curriculum by excercising on each other, by having many (simulated) patient contacts, by excercising on training models and by means of audiovisual aids.

The special emphasis on this aspect of the medical study is reflected in the evaluation system of students' achievements. Once a year all students (of all six curriculum years) are submitted to a so-called 'Skills Test'. The test resembles an Objective Structured Clinical Examination (OSCE, Harden & Gleeson, 1979) in that the tested skills have to be actually demonstrated. It also has circuits of stations through which students rotate. The test differs from an OSCE in the way scores of students are gathered. The method used in the Skills Test is fully restricted to a standardized observation method.

Students are brought into a controlled environment or standardized situation in which all circumstances are held constant. Their behavior while performing a specific skill is assessed by one (or more) observer(s) with the help of detailed checklists. Each skill is unraveled to its smallest parts and translated into operational terms, forming the items of the checklist. All checklists are constructed by a special committee. Observers, regular faculty members, are trained for each test so that ambiguities are banned and consensus on interpretations and conclusions is reached.

Every year separate tests are constructed for each curriculum year, reflecting the proceeding educational skills program. A more detailed description of the Skills Test and acquired experiences are to be found in Van Luyk, Van der Vleuten & Peet (1985). For an integrated delineation of the entire evaluation system this Skills Test is part of, see Verwijnen et al., (1982).

The study reported here will deal with validity aspects of the Skills Test. An experiment was carried out to investigate the test's concurrent validity and, if possible, to answer questions concerning its construct validity.

The validity problem

The assessment method of the Skills Test can be characterized as a purely analytical scoring procedure. Scoring of a student's competency is completely restricted to the dichotomous classification into behavior shown and not shown.

The intention is to minimize human inference, knowing that this causes variation within their own and between one another's conclusions (cf. Wiggins, 1973). This sometimes leads to comments by observers that their personal opinion with respect to the competency of a particular student is in conflict with the outcome of the checklist. Would this occur frequently then there would be a serious problem. Especially in the case where observers think

poorly of the examinee, but where the checklist scores show the opposite. Taking the observer's opinion as criterion, one could classify these kind of errors as false positives. Particularly in health care situations these false positive errors are not without danger, as it would mean that imcompetent doctors are certified.

In essence the above errors are directly connected with the validity of the test. The problem can be paraphrased as follows: How well does the test measure what it is supposed to measure and if we conclude that students are either poor or good performers, what is the quality of this conclusion?
More specifically, the following questions are to be put:

1. What agreement is there between outcomes of the checklist and the observer's judgement of student performance in a particular field?
2. If the outcomes show discrepancies, how can they be accounted for and explained?

The former question refers to the test's concurrent validity in that it is related to a criterion (observer opinion) obtained at the same time from the same behavior. The latter rather refers to construct validity.

Discrepancies between observer qualifications and test outcomes can originate from at least two sources. First, behavior scored by means of the checklist can be weighted differently by observers. It is possible that some aspects are given stronger positive, or, more realistically, negative weights by observers, yielding unequal final qualifications. The second plausible reason for disagreement may come from student performance not covered by the checklist. Best represented in the Skills Test are items with respect to technique or execution of a particular skill. Other aspects, such as attitudes or personal characteristics of examinees, are less represented. Especially the second question might shed light on the construct validity of the Skills Test: What aspects of skillful student behavior are essential and are reflected or neglected in the checklists used?

A final element of this study concerns a methodological issue. As has become clear above, two different methods of evaluation are used. The crucial disparity in the two methods is the way in which data are gathered.

Scoring by means of a checklist is simply checking and marking of behavior shown. In its strictest sense there can be no personal influence on the part of the person scoring. Like multiple choice items, checklists are unequivocal and leave less room for dispute about the scoring-procedure. This method can be called analytical. By contrast, judgements or ratings by observers, whether we try to make them more objective (e.g. Likert scales, anchors) or not, never exclude ambiguity. It is an established fact that if interpretations are needed, variations occurs and judgements become prone to halo-effects, leniency influences, or any other response-style. As opposed to the analytical method, this scoring procedure can be called a global evaluation method.

The incorporation of a global evaluation method in a study in order to validate an analytical procedure seems contradictory. There is, however, a fundamental difference between the global judgement procedure used here and the regular usage of global ratings. In general, global rating methods are often not applied in situations where target behavior is directly observed - as is done in this case - but where often reliance is placed on data from unstructured situations over a relatively long period of time (Fiske, 1971). Many examples

of the application of the global rating method can be found in clinical situations (cf. Katz & Snow, 1980; Morgan & Irby, 1978). The ratings in the experiment at hand are based on a short period of directly observed and specific behavior performed under standard conditions.

Method

In the academic year 1983-1984 observers in the Skills Test were asked to complete a 9-item rating scale in addition to their regular checklist. Items were in a Likert-format, consisting of a seven-point numerical scale, ranging from very poor to very good.

The first item asked for a general impression with respect to the skill performed. Here observers were able to give a total and general qualification of performance. The intention was to elicit a general opinion from the observer on the total achievement with respect to the skill performed.

The other items in the rating scale asked for specific elements of performance, which could be relevant for the total impression. The following specific items were used:

1. completeness
2. systematic action
3. efficiency
4. fluency
5. data-gathering
6. data-interpretation
7. self-confidence
8. patient-centeredness

The first four of these items reflect the quality of the skill-execution, items 5 and 6 cognitive or problem-solving aspects, item 7 a personal quality of the examinee and item 8 an attitude or the quality of the interaction process with the patient. When items were inappropriate, they were left blank as being 'non-applicable' response options.

Only those observers who already had experience with the Skills Test were approached, because they were supposed to have enough time left to fill in the rating scales. Although they were more or less expert observers, this may have introduced some sample bias. The observers were not specially trained to work with the rating scale. Every item on the list was briefly explained in a separate written instruction. Since the administration of these ratings took place within a regular (summative) Skills Test context, a requisite was that normal test procedures remained unaffected and that completion of the rating scales would be easy and of short duration.

The request led to 600 responses. A number of them were not correctly completed (missing or incorrect identification numbers) and therefore discarded and 566 ratings remained. They were approximately equally divided over the six curriculum years and over the various skill-domains (cf. Van Luyk, Van der

Vleuten & Peet, ibid.), with the exception of the so-called 'social skills' (interviewing techniques etc.). This latter skill-domain was itself part of a research program and, for lack of time, observers were not able to complete the global rating scale in these stations.

Observers completed the global rating scale after they completed the checklist. This inevitable condition might have had a considerable influence on the global ratings. Therefore two control experiments were carried out. These will be discussed later.

Results

General impression rating

Figure 1 shows the frequencies of the general impression rating, the first item of the rating scale, and the corresponding percentage correct scores on the checklists. These two sets of data will be compared in the first analysis.

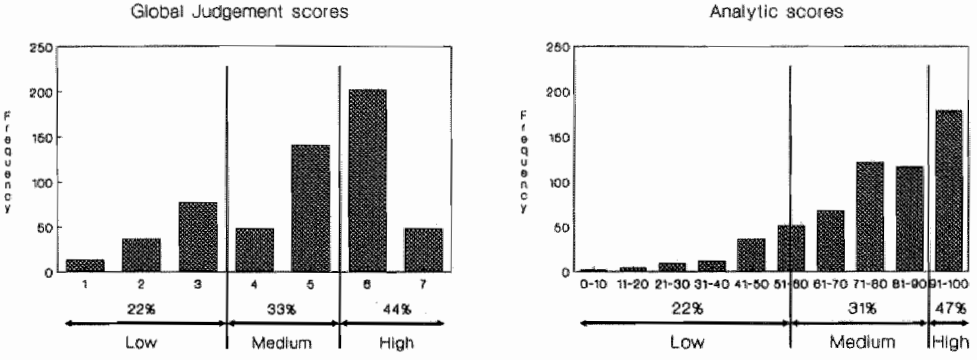

Figure 1: Frequencies of general impression ratings (global judgements) and corresponding station scores (analytic scores).

As can be seen in figure 1 the checklist scores are strongly negatively skewed. Although negatively skewed distributions are more often found (especially in the first few curriculum years) this distribution is not representative of regular score-distributions of stations. There is also some skewness in the general impression rating. The specific items in the rating scale (not shown) have similar shapes with peaks centering around the scale points 5 (fairly good) and 6 (good).

For further analysis it was decided to categorize the data into three areas. A lower area, corresponding with poor performance, a higher area referring to very good performance and an intermediate section representing mediocre performance. Cut-off points were chosen rather arbitrarily, with the restriction that

the frequency of cases within categories were about equal for both methods. These categories are also delineated in figure 1.

The question of agreement between global rating and checklist outcome can be dealt with by comparing the classifications of both evaluation methods. In table 1 a contingency table is given.

Table 1 shows, as was to be expected, the highest values on the diagonal elements and lower values on the off-diagonal cells. The hypothesis of no agreement must be rejected ($X^2 = 179.39$, df = 4, $p < .0001$).

Table 1: Cross-tabulation of global versus analytical judgement of technical and clinical skills.

		GLOBAL				GLOBAL		
		Low	Medium	High		Low	Medium	High
	Low	71	40	12	123	56%	21%	5%
Analytical	Medium	39	79	59	177	31%	42%	24%
	High	17	70	179	266	13%	37%	72%
		127	189	250	566	100%	100%	100%

More salient however is not the overall concordance between the two methods, but the locality of disagreement. Omitting misqualifications of adjacent categories, expecting similarity between these cells would be very severe, two significant cells are left to look at. The first is a high station score in combination with a low general impression rating: 17 cases out of 566 are to be counted in this cell. The second represents a low station score with a high general impression rating: 12 cases in this cell. On a total of 566 cases, these differences are not very impressive. The two ways of assessing performance yield a fair degree of convergence and only about 4 percent is qualified in the opposite way.

Analyzing the type of errors made, may provide a different view on the results. Taking the perspective of the observer and looking at checklist outcomes, our original standpoint, other inferences become possible.

A high analytical judgement of a examinee performance combined with a conflicting observer opinion was labeled, as a false positive error. Conversely, a high observer opinion in combination with a low station score was a false negative decision. The right part on table 1 is a re-ordering of the results according to the observer perspective. The contingency table is vertically percented, so horizontal comparisons can be made. The following inferences are possible:

- If students are rated as poor performers by observers, the chance of being qualified as a good performer by the checklist is 13 percent (false positives).

- If students are rated as good performers by observers, the chance of having a low checklist score is 5 percent (false negatives).

The conclusion to be drawn of course depends on the (subjective) weighting of magnitude of these errors. In general, false positive errors are considered more serious than false negatives, particularly with respect to clinical skills. A zero frequency of errors seems unrealistic, especially when the complexity of this testing-method is taken into account. On the other hand, the percentage of false positive errors is considerable indeed and is by no means to be neglected.

Specific ratings

In the above comparison, the global judgement procedure consisted of a single general rating. To further elaborate on the decision making errors and to trace their possible determinants, the other items of the rating scale may be relevant.

In table 2 frequencies are listed of the eight remaining items of the total global rating scale, referring to only one cell of table 1, namely the one containing cases of high checklist scores and low general impression ratings. In other words, for all false positive decisions the frequencies on the specific items are given in table 2.

Table 2: Frequencies of specific global rating items of students with a high analytic score and a low general impression rating.

	GLOBAL RATING						
	Absolute frequencies			Percentages			
Item	Low	Medium	High	Low	Medium	High	
Completeness	8	5	4	47	59	24	100
Systematic action	9	4	3	56	25	19	100
Efficiency	8	5	4	47	29	24	100
Fluency	9	6	0	68	40	0	100
Data-gathering	2	9	1	17	75	8	100
Data-interpretation	2	8	2	17	67	17	100
Self-confidence	10	7	0	59	41	0	100
Patient centeredness	8	6	1	53	40	7	100

On a priori grounds one may hypothesize that disagreement between observer opinion and checklist outcome is to be expected with respect to those areas that are least or not covered by the checklist. Checklists best represent technical and psychomotor aspects of performance. Next to these, there are items reflecting data-gathering, inferences made and conclusions drawn from them, but there are no items referring to attitudinal or personal characteristics of examinee performance. Particularly on these items examinees might have shown a poor performance possibly causing a low overal impression of performance. On items with regard more technical areas, a high degree of agreement with checklist outcomes could be expected. Close inspection of table 2 shows that this

hypothesis is only marginally true. Item 7, the impression of the student's self-confidence, answers this expectation most. More surprising is the fact that items 1 to 6 are not in congruence with the hypothesis. On the contrary, rather the opposite can be observed. Observers also appear to disagree on items well-represented in the checklist.

The same analysis can be undertaken for the other cell of table 1 referring to false negative errors. Table 3 shows these results.

Table 3: Frequencies of specific global rating items of students with a low analytic score and a high general impression rating.

	GLOBAL RATING						
	Absolute frequencies			Percentages			
Item	Low	Medium	High	Low	Medium	High	
Completeness	1	4	7	8	33	58	100
Systematic action	1	2	9	8	17	75	100
Efficiency	1	3	8	8	25	67	100
Fluency	0	3	6	0	33	67	100
Data-gathering	0	4	5	0	44	66	100
Data-interpretation	2	3	5	20	30	50	100
Self-confidence	1	4	7	8	33	58	100
Patient centeredness	2	6	3	18	55	27	100

Table 3 confirms even in a somewhat stronger way the paradoxal finding that observers are inclined to disagree on items well represented in the checklist. High frequencies are seen on items which are also, and possibly more objectively, represented in the checklist.

The unexpected locality of disagreement shown by tables 2 and 3 may also be accounted for by the fact that there may have been items influencing the general impression that were not included in the global rating scale. To explore this explanation a multiple regression analysis was conducted on the global ratings. The overall general impression was taken as a criterion and the specific items as predictors (see table 4).

The multiple R squared shows that 43 percent of the variance of the general impression rating can be accounted for by the specific items. This leaves 57 percent unaccounted. Probably other aspects are also determining the overall impression, which are inadequately or not covered at all in the list of items.

The regression analysis also provides estimations of the weights (beta) of the items, reflecting the importance of the separate aspects in relation to the general impression. Taking statistical significance as an indicator, completeness, data-gathering (surprisingly negative) and data-interpretation seem to be the decisive predictors for a general impression of an examinee performance. These items are also well covered by the checklist.

Table 4: Multiple regression analysis of specific global judgements with general impression as dependent variable.

Multiple R : 0.66
R^2: 0.43

Item	Weight (B)	Standard Error	Standardized Weight (Beta)
Completeness	5.14	0.84	0.41[1]
Systematic action	0.58	0.95	0.05
Efficiency	1.46	0.88	0.12
Fluency	0.06	0.99	0.01
Data-gathering	-2.69	1.02	-0.18[1]
Data-interpretation	3.40	1.01	0.25
Self-confidence	0.93	0.95	0.07
Patient centeredness	-0.46	0.70	-0.03

[1] $p < 0.01$

In summary, it may be concluded that there is a fair agreement between the general impression observers have of examinee performance and the checklist outcome. Taking the perspective of the observer, there are predominantly false positive errors and a smaller number of false negative errors. Closer inspection of these errors shows that observers are inclined (also) to disagree on matters well-represented (and probably well-assessed) in the checklist.

Control experiments

The above results may have been affected by the condition under which the global ratings were gathered. Administration took place within the regular context of the Skills Test and standard testing conditions had to be respected. As a consequence, global ratings were completed after observers filled in the checklist. The effect of this condition may have led to a considerable bias in the global rating by observers. Two control experiments were carried out to estimate the 'checklist bias'.

During the Skills Test administration video-tape recordings were made of randomly selected stations visited by fourth and fifth year students. From this material a sample was taken of 24 different stations, representative for the sample of stations in which the original global ratings were filled in. It was taken care of that almost all skill-domains were covered, except for the social skills (which were not represented in the original sample of global ratings either). With the help of these 24 tapes of students, recorded at different stations, two small generalizability experiments were carried out.

First, to estimate the influence of the completion of a checklist before filling in a global rating scale, three observers were asked to complete the global rating scale on two different occasions with the same (video-taped) students. On one occasion they were asked to first complete a checklist and subsequently

the rating scale. On the other, no checklist was given and solely the rating scale was completed. Between the two occasions there was a lapse of three weeks. To balance for learning effects one observer started with the checklist condition on the first occasion and the two others on the second occasion. The order of the video-taped students was randomized and not the same on both occasions. The observers were recently graduated physicians who received a small financial compensation for their efforts. In advance they were given the same training an observer normally receives in preparation for a Skills Test. Observers were kept unaware of the fact that they had to rejudge the same students the second time.

In summary, the experiment was so arranged that it yielded a completely crossed four-facet generalizability design: Students (Persons), Items, Judges and Condition (Cronbach, Gleser, Nanda & Rajaratnam, 1972). Students and Judges were considered as random facets, Items and Condition were considered fixed facets.

Variance associated with the condition facet can be conceived of as checklist bias, unless observers do not score consistently on different occasions. In other words, within the condition variance in the above experiment, intra-observer variance is confounded.

To estimate the strength of this effect a second experiment was carried out. The reason for opting for two experiments instead of for one, was that in a single integrated and completely crossed experiment an observer would have to judge the same student four times. Too strong learning effects were then to be expected. Except for the condition facet, this second experiment was completely identical to the first one. Another three observers were asked to judge the same video-taped students. This time on both occasions solely the global rating scale was used.

Tables 5 and 6 summarize the results of the two experiments.

The estimated variance attributable to the Condition facet yields a negative variance component indicating no direct influence of checklist scoring on global judgements. As a consequence there can be no strong (confounded) intraobserver variability. The results of the second control experiment confirms this conclusion. A very small percentage can be attributed to direct differences between the first and the second administration of the global ratings. The effect of recognance of course might be a plausible alternative explanation for this finding.

Both tables show a specific erroneous effect. The interaction term in table 5, condition by persons by judges, is relatively high. This means that some judges rated some examinees inconsistently on the two occasions. This effect is probably not influenced by the checklist bias, seeing that there is a remarkably identical effect in table 6. Here this variance component even explains almost 40 percent of the total variance. In conclusion, it can be stated that in these experiments the effect of checklist scoring on global ratings is negligible. Within the limitations of these experiments, this conclusion validates the earlier reported results on the comparison between checklist and observer qualifications of students.

Table 5: *Analysis of variance and estimates of variance components of global ratings under conditions either with or without checklist.*

	Mean Squares	df	Estimated Variance Component	Percentage of Total Variance
Persons (P)	40.47	23	0.6019	28.31
Items (I)	5.02	8	0.0056	0.26
Judges (J)	111.60	2	0.2377	11.18
Condition (C)	4.14	1	0.0000[1]	0.00
P x I	1.45	184	0.1590	7.48
P x J	7.13	46	0.1789	8.42
P x C	3.71	23	0.0000[1]	0.00
I x J	2.34	16	0.0324	1.53
I x C	1.62	8	0.0136	0.64
J x C	4.04	2	0.0000[1]	0.00
P x I x J	0.56	368	0.0407	1.91
P x I x C	0.42	184	0.0000[1]	0.00
P x J x C	3.83	46	0.3726	17.53
I x J x C	0.70	16	0.0094	0.44
P x I x J x C	0.47	368	0.4742	22.31

[1]Negative variance components set to zero.

Table 6: *Analysis of variance and estimates of variance components of global ratings with absent checklist scoring.*

	Mean Squares	df	Estimated Variance Component	Percentage of Total Variance
Persons (P)	14.91	23	0.2999	13.30
Items (I)	10.34	8	0.0560	2.49
Judges (J)	1.75	2	0.0000[1]	0.00
Condition (C)	9.70	1	0.0072	0.32
P x I	0.72	184	0.0849	3.98
P x J	4.30	46	0.0000[1]	0.00
P x C	2.49	23	0.0000[1]	0.00
I x J	0.80	16	0.0000[1]	0.00
I x C	2.37	8	0.0173	0.77
J x C	9.92	2	0.0037	0.16
P x I x J	0.50	368	0.0000[1]	0.00
P x I x C	0.59	184	0.0000[1]	0.00
P x J x C	8.58	46	0.8530	37.93
I x J x C	1.44	16	0.0224	1.00
P x I x J x C	0.90	368	0.9012	40.07

[1]Negative variance component set to zero

Discussion

Two aspects of validity of the Skills Test were dealt with in this study. The concurrent validity question of agreement between checklist outcomes and observer qualifications can be answered positively. In general, there was a fair degree of agreement between the two methods[2]. Taking the perspective from the observer, there was a rather high percentage of false positives and a smaller but not to be neglected percentage of false negatives. In the eye of the observer there were a greater number of students who should have failed, but who actually passed, and, conversely, a smaller amount of students who should have passed, but actually failed. Tracing the causes of these errors by comparing specific global ratings, yielded an unanticipated outcome. A priori it was assumed that if discrepancies would appear, they were likely to be found in areas marginally or not covered by the checklists, such as personal and attitudinal factors. This appeared only minimally to be the case. There was even a small tendency in the data showing the opposite: observers were inclined to disagree on aspects best represented in the checklist. In spite of this discordance, this unexpected result strengthens the concurrent validity of the Skills Test. Disagreement, unfavourable for the validity question, becomes supportive when this disagreement is on matters which are likely more objectively and validly measured by the Skills Test.

The multiple regression analysis of specific items on the overall impression rating, showed that a large portion of variance remained unexplained. Presumably, there are other elements which are also relevant for a general impression of skilled performance, not incorporated by the specific items used here.

The question of construct validity is a more complicated one. Especially the disagreement data were of importance for support of this validity aspect. As stated before, possible specific (in)validating areas may not have been found, since not all aspects of the general impression were covered by the specific items. On the other hand, on the aspects that were included, no considerable differences emerged: disagreement did not focus on elements least represented in the Skills Test. In conclusion, these outcomes may be considered as a confirmation of the construct validity of the Skills Test.

In addition, it can be argued that the aspects of skilled performance that are most validly assessed in the Skills Test are the quality of skills-execution, data-interpretation and data-management. These facets also appeared significant factors in the multiple regression analysis. Elements considered most important in skilled performance by (expert) observers converge with what is well-represented in the Skills Tests. So here too, we can infer some support for the construct validity of the Skills Test.

The control experiments clearly indicated that an obvious source of bias in the ratings global due to prior checklist scoring was negligable. Ratings with or

[2]The actual agreement is probably even higher. All analyses here were conducted at the station level. An examinee score on the total Skills Test is however a composite of multiple station scores. It is likely that the disagreement is "averaged out" when a composite score over multiple stations is calculated, yielding a smaller amount of disagreement overall.

without prior checklist scoring yielded comparable results. It should, however, be realized that in generalizability experiments like these with small sample numbers, the standard error of variance components is considerable. Nevertheless, these experiments unambiguously accredited the earlier findings, gathered during the Skills Test under sub-optimal conditions.

The outcomes of this validity study give us information about global ratings too. Instead of looking at the validity problem of the (analytical) Skills Test one might converse the point of view and validate the global ratings by considering the analytical procedure as criterion. Then of course it can also be concluded that there is a fair degree of agreement between this global rating procedure and the analytical scoring procedure. Questions can be posed with respect to the inter-rater reliability of global ratings (cf. Steiner, 1985) but it is interesting to note that global ratings may have a fair degree of validity. Once again, one should realize that there were special circumstances under which these ratings were gathered: a strongly structured standardized performance situation. Contrary to this study, judgements by means of rating scales are often characterized by unstructured situations, in which extensive generalizations have to be made on the basis of a small number of cues (Wiggins, 1973). Probably these conditions are responsible for the fair degree of validity the global evaluation method ascertained here. Further research will be needed to explore this validity.

References

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles.* New York: John Wiley.

Fiske, D.W. (1971) *Measuring the Concepts of Personality.* Chicago: Aldine Press.

Harden, R.M. & Gleeson, F.A. (1979) ASME Medical Education Booklet No. 8. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education.*

Katz, F.M. & Snow, R. (1980) *Assessing Health Workers' Performance: A Manual for Training and Supervision.* Geneva: World Health Organization.

Luyk, S.L. van, Vleuten, C.P.M. van der & Peet, D.G.M. (1985) The assessment of clinical and tecnical skills at the Maastricht School of Medicine. *Paper presented at the International Conference on Newer Developments in Assessing Clinical Competence, Ottawa, Canada.*

Morgan, M.K. & Irby, D.M. (1978) *Evaluating Clinical Competence in the Health Professions.* Saint Louis: C.V. Moss.

Schmidt, H.G. (1983) Problem based learning: Rationale and Description. *Medical Education, 17,* 11-16.

Verwijnen, G.M., Imbos, Tj., Snellen, H., Stalenhoef, B., Pollemans, M., Luyk, S. van, Sprooten, M., Leeuwen, Y. van & Vleuten, C. van der (1982) The evaluation system at the Medical School of Maastricht. *Assessment and Evaluation in Higher Education, 3,* 225-244.

Wakefield, J. (1985) Global rating scale. In: Neufeld, V.R. & Norman, G.R. (Eds.) *Assessing Clinical Competence.* New York: Springer.

Wiggins, J.S. (1973) *Personality and Prediction: Principals of Personality Assessment.* Reading: Addison-Wesley Publishing Company.

HOOFDSTUK 6

ASSESSMENT OF CLINICAL SKILLS WITH STANDARDIZED PATIENTS: STATE OF THE ART

Summary

A little more than ten years ago, the Objective Structured Clinical Examination (OSCE) was introduced. An OSCE includes a number of "stations", at which examinees are required to perform a variety of clinical tasks. Although the same OSCE may involve a range of testing methods, standardized patients (SPs), non-physicians trained to reproducibly play the role of a patient for assessment purposes, are used most frequently. SP-based tests have become very popular for assessment of clinical skills, and many medical schools have now introduced these examinations.

In the past few years, more than a dozen studies have investigated the psychometric characteristics of SP-based tests; this article provides a comprehensive review of this work. It is divided into five major sections. In the first, studies are discussed individually, including a description of the examinees, the station formats used, the test administration procedures, and the major psychometric findings. The second section discusses the reliability (reproducibility) of SP-based scores and pass/fail decisions, integrating results across studies through use of generalizability theory. The next section summarizes research on the validity of SP-based test scores. The fourth section discusses the impact of SP-based tests on the educational process. In an effort to address the practical needs of SP-based test users, the second, third, and fourth sections are structured as a series of responses to key questions in SP-based test design. The last section summarizes the state of the SP-based testing art, presents some ideas for improvement of SP-based tests, identifies several areas for further research, and provides some methodological observations and recommendations for future investigations of SP-based assessment techniques.

Across studies, reliability analyses consistently indicate that the major source of measurement error is variation in examinee performance from station to station, also known as "case-specificity" in the medical problem solving literature. As a consequence, it is necessary to include large numbers of stations in order to obtain a stable, reproducible assessment of examinee skills. Somewhat surprisingly, disagreements among raters observing examinee performance and differences between SPs playing the same patient role have much less effect on the precision of scores.

Results of conventional validation studies (e.g., differences in group performance; correlations with other measures) are generally favorable, though not particularly informative. Future validation research should include investigation of the impact of station format, timing, and instructions on examinee performance, in-depth study of the procedures used to translate examinee behaviour into station and test scores, and work on the potential impact of rater and SP bias.

Several recommendations are offered for improvement of SP-based tests. These include 1) focussing on assessment of history-taking, physical examination and communication skills, with written tests used to measure diagnostic and management skills, 2) adoption of a mastery-testing framework for score interpretation, perhaps using sequential testing methods, and 3) development of better standard setting procedures. Use of generalizability theory in analysis and reports of future psychometric research is also suggested.

Introduction

Medical schools and other organizations responsible for certifying clinical competence have traditionally made their judgments based upon written examinations and faculty ratings of performance in clinical training. In recent years, there has been a growing dissatisfaction with these procedures, because of the limited skills assessed by written tests and the psychometric problems associated with ratings of performance. This has led to a new emphasis in assessment, in which the tasks presented to examinees are more representative of those faced in real clinical situations. These performance-based tests are becoming very popular, and medical schools worldwide have begun using them for assessment of clinical skills (cf., Hart et al., 1986; Hart & Harden, 1987). Increasingly, schools are conducting and reporting studies of the psychometric characteristics of the tests they use, and a number of consistent results, both encouraging and discouraging, are evident. This article reviews the psychometric studies of one family of performance-based tests, those involving standardized patients.

Some Definitions and Terminological Distinctions

The OSCE. About fifteen years ago, the Objective Structured Clinical Examination (or OSCE) was introduced by Harden and colleagues (1975), though similarly structured "practical exams" and oral exams have been used in medical training for centuries. An OSCE involves examinees rotating around a circuit of stations at which they are required to perform a variety of clinical tasks. These tasks may include taking a brief history from a patient, performing some portion of a physical examination, undertaking an emergency procedure, or interpreting investigational data. The time allowed at each station can vary from a few minutes to an hour, depending upon the task to be performed, but, most commonly, stations last from five to twenty minutes. Examinee performance is scored on detailed checklists and rating forms tailored to the content of each station. Essential to the OSCE procedure is the combination of clinically relevant tasks tailored to the skills to be assessed, controlled, standardized testing situations, and predefined grading criteria. These are thought to provide a significant advantage over traditional unstructured and uncontrolled ratings of performance in clinical training. The OSCE is not really a testing method, however: it is a flexible approach to test administration, in which a variety of methods can be embedded in order to obtain a broad-based assessment of clinical skills. Individual OSCE stations may use any type of assessment method, varying from demonstration of procedural skills on cadavers to multiple choice questions. Most frequently, however, standardized patients (SPs) are used to test "hands-on" clinical skills.

Standardized Patients. SPs are non-physicians taught to portray patients in a standardized and consistent fashion for testing purposes. SPs can be asymptomatic, can have stable abnormal findings on physical examination (heart

murmurs, pulmonary crackles, joint abnormalities, etc.), or can simulate various physical findings (e.g., abnormal reflexes, diminished breath sounds, elevated blood pressure; different affects and personalities). Examinees can then interact with SPs as though they were interviewing, examining, and counseling real patients. Often, SPs are also trained to complete checklists and rating forms at the end of each encounter, recording the history information obtained, the examination maneuvers performed, and the counseling provided, as well as rating the communication skills of examinees. Alternatively, faculty-raters may observe SP-examinee encounters and complete checklists and rating forms.

Variation in Use of SPs. SP-based tests have been used to assess a broad range of clinical skills. Most often, they are used to measure history-taking, physical examination, and communication skills, though assessment of skills in diagnosis, laboratory utilization, and patient management are sometimes linked to SP stations using embedded oral examinations or written questions. A broad range of station formats has been developed; these vary in the skills assessed, the time required, the scoring criteria used, and other dimensions. For example, the classic OSCE station (Harden & Gleeson, 1979) is about five minutes in length, with examinees taking a brief history of a present illness or performing an isolated component of a physical examination. Other investigators have used very long stations, requiring examinees to perform a complete history and physical, followed by written questions regarding the case. Others use intermediate-length stations, with examinees asked to do a focussed history and physical. Depending upon the resources available and the preferences of the test developers, examinee performance may be scored by SPs, other non-physicians, or physician-observers. All stations included in an examination may be identical in form or may vary extensively, usually depending upon the range of skills that test developers thought appropriate to assess.

Stations: Building Blocks for Tests. All investigators using SP-based assessment use the term "station" to refer to the basic building block from which tests are constructed. However, some investigators use the term to refer to each discrete element in a test, even if the elements are related to the same presenting situation; others use the term to refer to a group of related elements. For example, a chest pain case, in which examinees take a history, read an electrocardiogram, provide a diagnosis, and initiate treatment, may be described as one to four stations, depending upon the preferences of the investigators. In this article, such multiple-element encounters are uniformly termed and treated as single stations, because the elements are not independent from test construction and analysis perspectives.

Station Scores and Subscores. The scores and scoring systems used by different investigator groups vary extensively. Some groups calculate only overall scores for each station, aggregating across the skills measured and the checklists, rating forms, and written test materials associated with a station; for these groups, our review of psychometric results must be based upon overall station scores. Other groups retain and use subscores for individual station components, most often by calculating one or more subscores based upon interaction

with an SP and one or more subscores related to written followup questions; an overall score for the station may or may not be calculated in addition. Because this review concerns SP-based assessment, when multiple subscores have been used, we have focussed on psychometric results for scores based upon direct interaction with SPs (generally those related to data-gathering and interviewing skills). Results for other scores are reported if they provide insight into key issues in test design and score interpretation.

Composite Test Scores and Subscores. Investigators also vary in the methods used to aggregate (sub)scores across stations: some groups form composite scores by averaging individual station scores, yielding a single composite test score; other groups form a composite score for each group of similar stations; a third alternative (used by groups calculating multiple subscores for each station) is to calculate a profile of composite scores corresponding to station subscores. Psychometric analyses may or may not be reported for all composite scores used, depending upon the purpose of the test, the skills viewed as important by the investigator group, the reliability of the subscores, and other factors. In accord with the primary objectives of the review, we have chosen to focus on composite scores based upon direct interaction with SPs. In reporting reliability results, we have adjusted total testing time and testing time per station to include only the time actually spent in interaction with SPs. The only exceptions occur in those studies where investigators reported only results for a single composite score that included both SP-based and written components; in these instances (explicitly noted in the text and tables), total testing time and testing time per station reflect both components. While this overall approach is somewhat artificial, because of variations in station format and test administration procedures, it was necessary in order to make cross-study comparisons more meaningful.

Organization of the Review

The remainder of the review is divided into five major sections. In the first, studies are discussed individually, including a description of the examinees, the station formats used, the test administration procedures, and the major psychometric findings. The discussions are based upon published and internal reports, supplemented by conversations with the investigators in some instances. The second section discusses the reproducibility (reliability) of SP-based scores and pass/fail decisions, integrating results across studies through use of generalizability theory (Cronbach et al., 1972; Brennan, 1983). The next section summarizes research on the validity of SP-based test scores. The fourth section discusses the impact of SP-based tests on the educational process. In an effort to address the practical needs of SP-based test *users*, the second, third, and fourth sections are structured as a series of responses to key questions in SP-based testing. The last section summarizes the state of the SP-based testing art, presents some ideas for improvement of SP-based tests, identifies several areas for further research, and provides some methodological observations and recommendations for future investigations of SP-based assessment techniques.

The diversity of SP-based stations and tests makes review of previous research difficult: in a very real sense, the SP-tests included in this review are each unique in the skills that were assessed and the station formats that were used. Nevertheless, SP-based tests from the same institution and investigator group usually share a number of common features. Consequently, for this review, studies have been grouped according to the institution(s) responsible for test development. If multiple studies were conducted at the same institution, they are discussed as separate "data sets" under the same institutional heading. If several publications emanated from the same data set, all are cited and integrated into a single discussion of the study design and results.

Criteria for Inclusion of Studies

To our knowledge, this review includes all published (and some unpublished) studies of the psychometric characteristics of SP-based tests in which four criteria are met. First, the test must have been administered to a minimum of forty medical students or residents. Second, all examinees must have been completed a minimum of three stations. Third, the total number of SP-examinee encounters (product of examinees and stations per examinee) must have been at least four hundred. While larger examinee and station sample sizes are clearly desirable for accurate estimation of key psychometric parameters, these minimum values were established in recognition of the logistical intricacies and substantial resource requirements of SP-based tests. Because the review integrates results across studies, we felt that fluctuations due to limited sample sizes would average out.

The last criterion for inclusion was that results of reliability and validity analyses were reported in sufficient depth to be interpretable. A number of studies were eliminated on this basis, because reports provided purely descriptive accounts without reliability or validity information, or because the information provided was based upon inappropriate estimation procedures. We revisit this problem at the end of the paper.

Review of Studies

University of Adelaide

Since 1979, the University of Adelaide in South Australia has administered a test battery with an SP-based component to graduating students as a part of a joint final examination in internal medicine and surgery (Newble, 1988). Psychometric analyses of the test battery, focussing on the SP-based component, were reported in Newble & Swanson (1988) and Swanson & Norcini (in press). The investigators pooled the results of the 1983 to 1986 test administrations, in which a total of 429 examinees were tested. The test battery consisted of two components: a Theory Test (one-hour multiple choice test) and a Clinical Test (1.5 hour short answer test; three to five SP-based stations; ten to twelve non-SP stations). SP-based stations were five (occasionally ten) minutes

in length. Each required students to perform a common clinical task (e.g., a portion of a physical examination, patient education, a diagnostic or therapeutic procedure, etc). Two days were required to test all examinees within a graduating class. A separate content-parallel test form was developed for each day of test administration. Each form was administered four times, testing twelve to fifteen students each time; examinee groups taking the test on the same day were kept apart for security reasons. Student performance was observed and scored by two clinical faculty using checklists tailored to station content and skills assessed (cf., Newble et al., 1978; Newble et al., 1980). There was some rotation of SPs and observers assigned to a station across and within test forms.

Swanson & Norcini (in press) reported median inter-rater reliabilities of 0.78 for ten minute physical examination stations, 0.68 for five minute physical examination stations, 0.50 for five minute patient education stations, and 0.76 for procedural skills stations (intraclass correlations with differences between raters included in measurement error). For generalizability and validity analyses, the investigators pooled results across years and test forms by calculating average variance components and covariances (Newble & Swanson, 1988). Use of variance components allowed calculation of a broad range of reliability indices for a variety of test lengths, conditions of test administration, and norm-/domain-referenced score interpretation. A generalizability coefficient (analogous conceptually to coefficient alpha) of 0.31 was reported for the SP component of the exam (five stations, half hour of testing time, two raters per station). Observed correlations between the SP-based test and other test components varied from 0.33 to 0.40, reflecting, in part, attenuation due to low reliability of the components. The investigators statistically disattenuated the observed correlations to estimate true correlations (the values which would be obtained if all tests were very long and perfectly reliable), using the generalizability theory analog of the Spearman-Brown prophecy formula. True correlations were quite high, ranging from 0.68 (SP stations and multiple choice test) to 1.00 (SP and non-SP stations). In general, true correlations between components of the Clinical Test were higher than those with the Theory Test.

College of Family Physicians of Canada

For more than a decade, the certification examination of the College of Family Physicians of Canada has included a Simulated Office Oral (SOO) component consisting of five clinical simulations similar to situations which family physicians commonly encounter in practice (Rainsberry et al., 1987a). In each SOO, an examinee has fifteen minutes to take a history from and/or counsel a patient role-played by an examiner who is also a College member and certified family physician[1]. The examiner scores examinee performance on case-specific

[1]Although the SOO is an oral examination, it is included in this review because the patient-role of the examiner is conducted in a very strict sense. Hence, the difference between a (physician) examiner as a patient, and the patient as an examiner (such as is the case in a number of studies below) is vague. More detailed information concerning the oral examinations of the College of Family Physicians of Canada can be found in a series of internal reports available from the College (Grava-Gubins et al., 1985a, 1985b, 1985c, 1986, 1988; Rainsberry et al., 1985, 1987b; Khan et al., 1988).

rating scales covering several skill areas (typically problem definition, affective skills management, knowledge of the family, practice organization, professional responsibility, and health maintenance); scales include criteria which operationally define failing, passing and excellent performance. Examiners are provided with extensive training, both in playing the patient role and in scoring performance.

Rainsberry et al. (1987a) reports the results of a series of small inter-rater reliability studies conducted during training sessions at multiple testing centers in 1984 to 1986. Intraclass correlations for the reliability of a single rating (mean differences between raters *excluded* from measurement error) ranged from 0.49 to 0.82, with a median of 0.63. Estimates of the overall reproducibility of scores (reflecting measurement error associated with both cases and raters) can be approximated from intercorrelations among SOO scores reported in factor analytic studies of 1984 to 1986 exams (Grava-Gubins et al., 1987); the scores of 1545 examinees, tested at centers across Canada, were included in these analyses. The resulting estimate for an exam including five cases is 0.59[2]. Relationships with other measures have not been reported.

Educational Commission for Foreign Medical Graduates

For the past several years, the Educational Commission for Foreign Medical Graduates (ECFMG) has worked on a Clinical Skills Assessment (CSA) Project, seeking to construct an examination which could be given to foreign medical graduates prior to entry to postgraduate training in the United States (Conn, 1986). The current version of the CSA consists of a pictorial multiple choice test and three cases; the latter include an SP-like component. Each case begins by providing an examinee with a brief written history, which he/she has three minutes to read (Conn & Cody, under editorial review). Next, an examinee has fifteen minutes (thirty minutes for one case) to perform a focussed physical examination on a normal adult who does not attempt to simulate the physical findings for the case (hence the term "SP-like"). During the physical examination, an examinee states the maneuvers done and the reasons for them; an observing physician scores the maneuvers and reasons on a case-specific checklist. Next, the examinee obtains a history in patient management problem (PMP) format using a latent image marker. Last, the examinee is provided with the abnormal physical findings for the case and manages the patient by responding to a series of PMP sections and multiple choice questions.

Two pilot tests of the CSA have been conducted. In the first, 181 volunteer examinees (117 ECFMG-certified foreign medical graduates, 64 graduates from U.S. medical schools) were tested by 92 examiners at five test administration sites. Large and statistically significant differences favoring U.S. medical graduates were obtained, but little psychometric information was reported (Conn, 1986). The second pilot test was larger scale in scale, involving 635

[2]Data were pooled across years of test administration by calculating the average (inter-case) correlation between SOOs from Tables 3, 5, and 7 in Grava-Gubins et al. (1987). This provides a reasonably accurate estimate of the reliability of scores on a single SOO, with mean differences between SOOs and examiners excluded from measurement error.

foreign medical graduates and 123 U.S. medical graduates (Conn and Cody, under editorial review) at 21 test administration centers nationwide. As in the first pilot test, foreign medical graduates scored markedly lower than the U.S. graduates. In supplemental analyses (Cody, 1988), the reliability of the total CSA score was reported to be 0.73; the comparable value for the SP-like component was 0.51[3]. Low positive correlations (0.17 to 0.37) were observed among exam components. An association between performance on the CSA and a previous written exam required for entry into graduate medical education was also noted.

University of Limburg

Beginning in 1982, the University of Limburg Medical School, in Maastricht, The Netherlands, annually administered an OSCE-format exam termed the Skills Test to students in all six years of training (Van Luyk et al., 1986). The focus of assessment varies across years of training, from simple interviewing and isolated technical skills in early years to more complex interviewing and physical examination skills in later years. In all years of training, the test is two hours long and includes six to eleven stations lasting ten, twenty or thirty minutes (fifteen minute average). Stations focus exclusively on "hands-on" skills and contain no written components, though SPs are not used at all stations. Trained faculty-observers score examinee performance on detailed checklists of behaviourally oriented criteria. In approximately 20% of stations, co-observers independently score examinee performance. To test one class of 150 students, two days are needed. Generally, four different test forms are constructed for security reasons, and each test form is replicated (same "circuit" of stations, different SPs and observers) four times to allow concurrent assessment of more examinees. Within a replication, the same SPs and observers are used for all examinees.

Van der Vleuten et al. (1988) reported the results of reliability analyses of all Skills Tests administered from 1984 to 1987. Only the results of Skills Tests given to fifth and sixth year students (544 examinees averaging eight stations each) are reviewed here, because stations in these exams are almost exclusively SP-based. Inter-rater reliability was reported to be 0.79 (intraclass correlation with differences between raters treated as measurement error), based upon scores from 913 stations where observers and co-observers were present. To estimate test reliability, variance components were calculated for each test form and then averaged across forms and calendar years[4]. Variance components and reliability coefficients are reported in the paper for a variety of test lengths and

---

[3]These reliabilities may be somewhat inflated, because different examiners rated different examinees. Consequently, differences between examinees may partly be the result of differences in scoring by the examiners.

[4]This approach ignored replications of circuits within test forms. This may have inflated the variance components associated with examinees and stations, since examinees and stations are partially confounded with SPs and rater groups within replications. The investigators noted that the impact on accuracy of variance components estimates was likely to be negligible, because inter-rater reliability was high and multiple raters were used within replications.

for norm-/domain-referenced score interpretation. The generalizability of a two-hour Skills Test like those typically given was estimated to be 0.69. Van der Vleuten et al., (1989) reports correlations of 0.63 (observed) and 0.77 (disattenuated) between scores on the Skills Test and a locally developed written test of general medical knowledge. Approximately the same correlations were observed between Skills Test scores and a written test of technical and clinical skills intended to reflect the cognitive component of the Skills Test.

University of Massachusetts/New England Consortium

In collaboration with consortia of medical schools and residency training programs in the New England region, investigators from the University of Massachusetts and the American Board of Internal Medicine (ABIM) have conducted three large scale studies of the psychometric characteristics of SP-based tests. The first data set discussed below derives from a 1984-85 study involving residents from fourteen internal medicine training programs. The second resulted from a study conducted in 1986 involving senior students at the University of Massachusetts, and the third from a 1987 study of senior students from five medical schools.

Data Set 1

Examinees in this study (Stillman et al., 1986a, 1986b) included 112 first-year residents, 92 second-year residents, and 132 third-year residents. All stations involved workup of common ambulatory problems in internal medicine. Each examinee was tested with four (occasionally three) stations from a pool of twelve cases. All stations were forty minutes long: examinees had thirty minutes to perform a focussed history and physical with a SP and ten minutes to answer a series of open-ended questions regarding differential diagnosis and initial management. SPs recorded the findings obtained by examinees on a case-specific checklist and scored interviewing skills on a rating form while examinees completed the open-ended questions. SPs traveled to the clinical facilities of the participating residency programs for test administration. Data collection required twelve months, and the investigators reported that "rough comparability" was maintained in the mix of residents working up each case. As a part of SP training, all stations were pilot tested and co-scored by an observing staff member prior to actual use. During data collection, the SP trainer occasionally observed encounters on a rotating basis and co-scored examinee performance. Various other measures (written test scores, ratings of resident performance from program directors and faculty, self-ratings) of examinee skills were also obtained.

Inter-rater reliability based upon observation during pilot testing (48 pairs of scores) was reported as 0.70 for checklists of historical findings, 0.70 for checklists of physical examination maneuvers, and 0.52 for the interviewing skills rating form (all product moment correlations). Based upon data collected during test administration (37 pairs of scores), these reliabilities were 0.82, 0.86, and 0.67, respectively. No inter-rater reliabilities for the written component were given. The investigators also reported "inter-case" generalizability coefficients (intraclass correlations with mean differences between stations/SPs

excluded from measurement error) for a variety of scores as a function of test length: for the data-gathering score (history and physical examination checklists combined), a coefficient of 0.74 was obtained for a test consisting of four stations, the test length actually used; the comparable value for interviewing skills was 0.52. The reliability of all scores related to differential diagnosis and laboratory utilization was substantially lower, leading the investigators to suggest that traditional written testing methods be used to assess these skills, with SP-based tests focussing upon history taking, physical examination, and interviewing skills. Scores were observed to increase with training, and residents from stronger programs performed better than residents from weaker ones. The investigators reported only statistically significant correlations, uncorrected for attenuation. For the data-gathering score, significant observed correlations were found with months of residency training (0.32), faculty ratings from one (of five studied) residency program (0.28), and the national certifying examination given at the successful completion of training (0.23).

Data Set 2

Examinees for this study were 96 medical students from the University of Massachusetts entering their final year of training (Stillman et al., 1987). The test consisted of fourteen SP-based stations of fifteen minutes each and a 1.75 hour written examination. Two types of SP stations were used: History Stations, where examinees had ten minutes to take a focussed history and five minutes to record a differential diagnosis and management plan, and Communication Stations, where examinees had five minutes to review background medical information related to the station and ten minutes to provide patient education and/or counseling. All stations involved common patient situations which graduating medical students should be able to manage; SPs scored examinee performance on checklists and rating forms as in Data Set 1. The written test consisted of open-ended questions and multiple choice items assessing primarily skills in interpretation of diagnostic studies (electrocardiograms, X-rays, blood smears, urinalyses, blood chemistries). Test administration took place over a four-week period, with students assessed in groups of ten. As in the previous study, all stations were pilot tested and co-scored by an observing staff member prior to actual use. Clerkship grades and scores on the National Board of Medical Examiners (NBME) Part I and II Examinations (the first and second components of the national licensure examination sequence) were retrieved from school records as traditional indicators of examinee proficiency.

While the mean level of examinee performance fluctuated from day to day, no consistent trends in performance were observed over the four weeks of test administration, leading the authors to conclude that security problems had not been serious. Based upon observation during pilot testing, inter-rater reliability was 0.93 and 0.77 for data-gathering and interviewing skills (product moment correlations), respectively. Inter-case reliability (generalizability) coefficients were reported for a variety scores as a function of test length. For the fourteen station test actually given, the reliability of the SP-based composite score, including both History and Communication Stations, was reported to be 0.78; reliabilities for scores related to differential diagnosis and laboratory utilization

(derived from History Stations only) were again very low. Observed correlations between the SP-based composite score and NBME Parts I and II were 0.19 and 0.27, respectively; SP-based scores correlated 0.44 with clerkship grades. The observed correlation between SP-based scores and performance on the concurrently administered written exam of interpretation skills was 0.26.

Data Set 3

In 1987, five New England medical schools (University of Massachusetts, Boston University, Brown University, University of Connecticut, Tufts University) collaborated on an assessment of the clinical skills of the fourth year students (Stillman et al., under editorial review). A committee of faculty representatives from the five schools developed an eighteen station SP-based test involving common clinical situations in several specialty areas. Three types of stations were used: twenty minute History/Physical Stations in which examinees had fifteen minutes to perform a focussed history and physical and five minutes to answer two questions regarding the patient's physical findings and differential diagnosis plus some multiple choice questions unrelated to the case; and fifteen minute History and Communication Stations similar to those described for Data Set 2. Each examinee took nine History/Physical, five History, and four Communication Stations. Test administration was also similar to Data Set 2: ten examinees were tested per day, with different, but content-parallel test forms used for security reasons. SPs recorded findings obtained by examinees on checklists and scored interviewing skills on rating forms as in previous studies.

Analysis of trends in examinee performance over the period of test administration yielded no evidence of security problems. Inter-rater reliability was not studied. Swanson & Norcini (in press), reports the results of generalizability analyses of a large subset of History/Physical Stations in which two SPs played each case role. This approach allowed estimation of a broad range of generalizability coefficients for various test lengths, conditions of test administration, and both norm- and domain-referenced score interpretation. For a three hour, twelve station test (single test form, single SP playing each role, norm-referenced score interpretation), generalizability coefficients of 0.67 and 0.85 were obtained for data-gathering and interviewing skills scores, respectively. Significant differences in performance were observed for groups tested by different SPs playing the same case role; this result was particularly pronounced for ratings of interviewing skills. However, for tests including several stations and random assignment of SPs to examinees, it appeared that SP-related differences in performance would not have a major impact on test scores. Scores for data-gathering and interviewing skills had observed correlations of 0.10 and -0.07 with NBME Part I, 0.22 and 0.05 with NBME Part II, 0.25 and 0.26 with clerkship ratings, and 0.08 and 0.00 with a self-assessment of clinical competence (Stillman et al., under editorial review).

National Board of Medical Examiners

In the mid-70s, the National Board of Medical Examiners (NBME) conducted a series of studies of methods to assess the interpersonal skills of medical students. One of these studies used SPs to measure the history-taking and

communication skills of forty fourth year medical students from the five medical schools in the Philadelphia area (Templeton et al., 1978). The test consisted of twelve SP stations (involving typical clinical problems which might be seen by fourth year students). At each station, examinees had twenty minutes to interview the SP, review a written summary of physical findings, and initiate treatment. Test administration took place over a two month period, with the same twelve cases and SPs used throughout; two half-days were required to test each examinee. All encounters were videotaped and scored by two trained observers using checklists of thirty to fifty "case-specific outcomes" (medical and psychosocial information which should be obtained; medical and psychotherapeutic actions which should be taken; interpersonal behaviours aiding in communication).

Inter-observer agreement was very good (average kappa coefficients of 0.78 across eleven checklists coded by both observers). Analyses indicated that the generalizability of a composite score based upon twelve stations (four hours of testing time) was 0.82. Correlations with NBME Part I and II composite scores and most subtest scores were non-significant; correlations of 0.37 and 0.32 were observed with Part I Behavioral Science and Part II Psychiatry subtest scores, respectively.

Southern Illinois University

In 1986, Southern Illinois University (SIU) began administration of an SP-based comprehensive examination early in the fourth year of medical school. Starting in 1987, the exam was developed jointly by faculty from SIU and the University of Manitoba, with the schools separately training SPs and testing their students. Data Set 1 is based upon the results of the 1986 and 1987 test administrations at SIU; Data Set 2 resulted from merging the results of the 1987 test administrations at SIU and Manitoba.

Data Set 1

A series of publications have described the 1986 and 1987 SIU Comprehensive Examinations included in this data set and reported the results of psychometric studies (Williams et al., 1987; Barrows et al., 1987; Williams & Barrows, 1987; Williams and Colliver, 1987; Dawson-Saunders et al., 1987; Colliver et al., 1989). The 1986 and 1987 exams included thirteen and seventeen SP-based stations[5] (plus several non-SP-based stations) taken by 70 and 67 fourth year students, respectively, six months prior to graduation. SP-based stations involved common clinical problems and were thirty to forty minutes in length. During the first half of this period, examinees performed a focussed history and physical examination and counseled SPs regarding treatment; during the second half, examinees responded to written questions regarding patient findings, diagnosis, and management. Test administration required approximately three

[5]Descriptions of the SIU exam in the literature are not completely consistent about the number of stations included and the testing time per station, probably resulting from different treatment of non-SP-based stations. Reliability results reported here were taken from Colliver et al. (in press), because it provided the most detailed information.

weeks; students were tested in groups of ten to fifteen, with roughly three days required to complete assessment of each group. The same cases were used throughout; two SPs were trained to play most case roles. Data-gathering and interviewing skills were generally scored by SPs on checklists and rating forms; written followup questions were scored by non-physician faculty using scoring keys developed by station authors. Clerkship grades and scores on the National Board of Medical Examiners (NBME) Part I and II Examinations (the first and second components of the national licensure examination sequence) were retrieved from school records for comparison purposes.

Analysis of trends over the period of test administration indicated that examinee groups did not differ significantly in performance, leading investigators to conclude that security breaches, if they occurred, had little influence on scores. Inter-rater reliability was reported for the 1986 data set only, on the basis of 40 encounters scored by SPs and an observer; 80% agreement was observed. Generalizability coefficients of 0.72 and 0.62 were obtained for station scores (SP-based and written components combined; values for SP-based scores were not separately reported) on the 1986 and 1987 exams, respectively. Colliver et al. (1989) provides an excellent summary of the results of generalizability analyses, both for scores and pass/fail decisions. Observed correlations between scores on the 1986 exam and other measures of clinical competence were 0.65 (clinical ratings), 0.53 (NBME Part I), and 0.51 (NBME Part II); these reflect performance on both SP- and non-SP-based exam components.

Data Set 2

The University of Manitoba administered 15 stations from the SIU 1987 exam to 67 volunteering senior students from their medical school (Klass et al., 1987); fifty-two students completed all stations. Written descriptions of the stations, videotapes of SIU SPs, and consultation with SIU staff provided a basis for training a new group of Canadian SPs by Manitoba collaborators. Test administration procedures were identical to those at SIU, except for use of different SPs. However, Manitoba students were unfamiliar with the testing format, and scores on the test were not used for grading purposes, while SIU students were very familiar with SPs and, in 1987, students had to achieve a passing score on the exam in order to graduate.

The investigators did not report information concerning test reliability; analyses focussed on comparison of performance at the two schools. Station means and standard deviations were very similar at the two schools, despite substantially different curricula, educational philosophies, and health care systems. The observed correlations of the exam with clinical ratings and NBME Part II were 0.52 and 0.63, respectively: these values are very similar to those obtained at SIU. The authors concluded that the exam was transportable across local and national boundaries.

University of Texas Medical Branch at Galveston

Researchers at the University of Texas Medical Branch at Galveston (UTMB) have conducted several studies of psychometric characteristics of SP-based tests, using data from assessments of clerks and residents in internal medicine. The first two data sets are derived from these sources. The last is based upon a

multi-institutional study; it is discussed in this section because UTMB was involved, and the testing procedures used were based upon previous UTMB work.

Data Set 1

The first data set was derived from administration of SP-based tests at the end of medicine clerkships during the 1983-84 academic year (Petrusa et al., 1987a; Petrusa et al., 1986; Petrusa et al., 1984). Sixty-eight junior students participated in each of three test administrations; in all, 204 examinees were tested. On average, tests consisted of 17 stations. Each station had two five minute components: an SP-component in which examinees performed some portion of a history or physical examination, and a written component in which examinees responded to short answer questions regarding physical findings, differential diagnosis, and management plan. Performance on the SP-component was scored on checklists by SPs; performance on written components was scored by non-physician faculty using scoring keys. Inter-rater reliability was investigated on both components by having physician faculty independently score performance at some stations. Thirty-four examinees were tested in a four hour period in the morning; the test was repeated in the afternoon for the 34 additional examinees. It is unclear whether the same 17 SPs were used for all 68 examinees.

Inter-rater reliability for the SP-based component averaged 0.80 (Cohen's kappa), and a similar value was obtained for inter-scorer agreement for the written component (product moment correlation). Test reliabilities (coefficient alpha) of 0.46, 0.49, and 0.57 were reported for the three test administrations. Performance in the last clerkship was best; this was interpreted as validity evidence, though it appears that test difficulty was confounded with group performance. Correlations with other measures showed moderate to high values: 0.46 with faculty ratings (0.73 corrected for attenuation) and 0.43 with the NBME Medicine Subject Examination (0.64 corrected). Scores on the SP-component weremore highly correlated with faculty ratings; scores on the written component were more highly correlated with the NBME examination.

Data Set 2

Examinees in this study were 95 first and second year residents in internal medicine (Petrusa et al., 1987b). They took the same examination as junior students in Data Set 1. The investigators estimated (internal consistency) reliabilities by pooling data across test administrations for nine consistently--used cases taken by 74 examinees; a reliability coefficient of 0.26 resulted. Both groups of residents (60 first year; 14 second year) performed significantly better than the student group, but there was no significant difference between resident groups. A significant, but low observed correlation was found with months of training, indicating a weak relationship between time in training and test performance. There was no correlation between test scores and supervisor ratings, but positive correlations were obtained for subscores related to interviewing skills (0.38 and 0.39, uncorrected for attenuation). Correlations with performance on one best answer, multiple true/false and patient management problem subscores of the ABIM Certifying Examination were 0.37, 0.24 and

0.31, respectively (0.54, 0.35, 0.51 disattenuated), indicating low to moderate relationships with written measures of competence.

Data Set 3

This data set was derived from a collaborative study conducted by UTMB, the University of Washington (UW), Southern Illinois University (SIU) and the University of North Carolina (UNC) (Petrusa, 1988). A team of investigators including representatives from all schools constructed a ten station SP-based test. Each school trained their own SPs to play the same station roles and then administered the test to junior students at the end of medicine clerkships. At some collaborating schools, faculty or residents observed and scored examinee performance; others trained SPs for this purpose.

Data was pooled across test administrations within site to estimate reproducibility of scores. The reported generalizability coefficients were 0.47 for UTMB (149 examinees), 0.50 for UW (44 examinees), 0.44 for UNC (92 examinees) and 0.25 for SIU (62 examinees). Schools differed significantly in performance; the patterns were reported to be consistent with differences in scores on the Medical College Admission Test. However, testing conditions varied across schools, making interpretation of school differences problematic. Test performance did not covary systematically with date of test administration: examinees taking the test later in the academic year did not perform better than those taking it earlier. Observed correlations with NBME Part I scores and clinical ratings were 0.64 and 0.37, respectively.

University of Toronto

In 1987, as a part of the process for selecting foreign medical graduates for Ontario clerkships, the University of Toronto developed a two-phase testing program (Cohen et al., 1988; Cohen et al., 1987). The first phase consisted of a multiple choice test administered to 233 examinees. In the second phase, the 71 examinees with the highest scores took an SP-based test consisting of 30 ten minute stations. At 21 stations, examinees had five minutes to take a history or perform some portion of a physical examination, followed by five minutes to respond to case-related short answer questions. Interviewing skills were tested at five of the remaining stations; the other four stations involved written test materials exclusively. At all SP-based stations, two faculty-raters were present; they alternated in scoring examinee performance on station-specific checklists. Two SPs were also used for each station; they also alternated. Examinees were randomly divided into two groups, with one taking the first half of the examination in the morning and the other in the afternoon; the second half of the examination was administered similarly on the following day.

Generalizability analyses indicated that the reliability of the composite score formed from SP-based station components was 0.83. The analogous value for the short answer component was 0.69, and total test reliability (SP-based and short answer components combined) was 0.86. Performance on the MCQ test given in phase one correlated 0.43 with scores on the SP-based component, 0.46 with scores on the short answer component and 0.48 with total test scores (all values uncorrected for attenuation). Analyses comparing ratings from faculty observers paired at the same stations indicated significant differences

between pairs on five of twenty-five stations; no significant differences were observed between scores derived from different SPs playing the same patient role on any of the twenty-five stations.

Reproducibility of SP-Based Scores and Pass/Fail Decisions

The previous section reviewed the results of a series of studies of the psychometric characteristics of SP-based assessment; Appendix 1 provides a summary of the examinees and tests used in these studies. This section discusses issues related to the reproducibility of SP-based tests, integrating results across the studies discussed individually in the previous section. We begin by presenting a general conceptual framework for thinking about the reproducibility of SP-based tests. This is followed by discussion of various factors influencing the reproducibility of SP-based scores and pass/fail decisions. This discussion is structured as a series of responses to key questions that commonly arise in designing and using SP-based testing procedures.

A Conceptual Framework for the Reproducibility of SP-Based Tests

The purpose of any assessment is to draw inferences about the ability of examinees -- inferences which extend beyond the particular sample of items included on the test to the larger domain from which the items are sampled. Depending upon the size and nature of the sample, estimates can be more or less reproducible (reliable) and more or less accurate (valid). Thus, test design can be viewed as the development of a sampling plan for the skills and areas to be tested.

For SP-based tests, SPs, raters (either faculty-observers or the SPs themselves), and stations are sampled from larger domains of SPs, raters and stations which might have been used on the test. In this context, reproducibility can be conceptualized as the extent to which an examinee's score would be stable across different but similar (randomly parallel) samples of SPs, raters, and stations. Reproducibility (generalizability) coefficients are best thought of as the expected correlation between scores derived from successive samples. For an estimate of an examinee's ability to be reproducible (e.g., a reproducibility coefficient greater then 0.8), an adequate number of SPs, raters, and stations must be included in the sample that the test comprises. Lack of inter-rater agreement in scoring examinee behaviour, inconsistency in SP performance, and variation in examinee performance across stations all affect the reproducibility of scores.

The structure of the second, third, and fourth subsections follows from this conceptualization of the reproducibility of SP-based tests. First, rater-related sources of measurement error are examined; this is followed by discussion of SP-related sources of measurement error. While SPs often act as raters as well as patients, it is important to keep the two sources distinct conceptually, because it is common practice to train multiple SPs to play the same role. Ratings (whether provided by an SP or an observer) could be perfectly accurate, with different SPs still varying extensively in how they play the same

patient role). In the fourth subsection, entitled "Station-Related Sources of Measurement Error," we discuss the impact of variation in an examinee's performance from station to station (referred to as *content-specificity* in the medical problem solving literature) on reproducibility of scores. This is the largest source of measurement error in SP-based tests; relatively long tests, including many stations, are required as a consequence in order to obtain reproducible scores.

Most psychometric analyses to date have (often implicitly) adopted a *norm-referenced* framework for score interpretation. The fifth subsection examines the impact on reproducibility of working within a *domain-referenced* testing framework. This approach is conceptually appealing, since advocates of SP-based testing often wish to interpret quality of performance in an absolute, real-world sense, rather than in terms of the relative ranking of examinees.

These subsections all focus on the *reproducibility of scores*. In the sixth subsection, we explore the adoption of a mastery testing strategy, in which the *reproducibility of pass/fail decisions* is of primary importance. As discussed in that subsection, the mastery testing approach can lead to a sizable reduction in testing time requirements, since making reproducible decisions generally requires substantially less precision than obtaining reproducible scores.

In large scale SP-based testing, it is necessary to develop a number of test forms for use at multiple sites over an extended period of time. These forms will generally differ in difficulty and discrimination; consequently, the score received by any particular examinee is influenced by the test form used. In the last subsection, we discuss the problem of statistically adjusting (*equating*) scores on alternate test forms to put them on the same scale.

Rater-Related Sources of Measurement Error

How well do raters agree in scoring individual SP stations?

Table 1 summarizes the results of analyses of inter-rater agreement for all studies reporting them. While the particular agreement index used varies from study to study, agreement is generally good. Similar results have been reported by Andrew (1977), Stillman et al. (1980), and Neufeld et al. (1983).

The values in Table 1 indicate inter-rater agreement *for individual stations*, not for the test as a whole. Because examinees are rated by different observers at each station, any measurement error introduced by inter-rater disagreement at individual stations will tend to average out across stations, as long as errors are non-systematically related to examinees (e.g., not associated with where the site where the exam was taken, not related to the examinees race, gender, or appearance, etc.). This issue will be explored further in later subsections.

How many raters are required per station?

Since inter-rater agreement is fairly good, it is unnecessary to use more than one rater per station, particularly for relatively long tests involving many stations. Swanson & Norcini (in press), in analyses of the University of Adelaide Data Set, provides a good quantitative illustration. In this data set, two physicians rated examinees at most stations. Table 2 presents the results of

Table 1: Inter-rater reliability of SP-based scores.

Data Set	Rater	Composite of	Value[1]
Adelaide	Faculty	10-min Physical Exam Checklists	0.74[2]
		5-min Physical Exam Checklists	0.68[2]
		Patient Education Checklists	0.50[2]
		Procedural Skills	0.76[2]
CFPC	Physicians	Ratings of Communication Skills	0.63[6]
Limburg	Faculty	Miscellaneous Checklists	0.79[2]
UMass Data Set 1	SPs	History Checklists	0.76[3]
		Physical Exam Checklists	0.78[3]
		Ratings of Communication Skills	0.60[3]
UMass Data Set 2	SPs	History Checklists	0.93
		Ratings of Communication Skills	0.77
NBME	Non-Physician Observers	History and Patient Management Checklists	0.78[5]
SIU Data Set 1	SPs	History & Physical Exam Checklists	0.80[4]
UTMB Data Set 1	SPs	History & Physical Exam Checklists	0.80[5]

[1]All entries are Pearson product moment correlations, unless otherwise noted
[2]Intraclass correlation with mean differences between raters included in measurement error
[3]Average between training session and test administration session
[4]Proportion agreement
[5]Cohen's kappa
[6]Intraclass correlation, mean differences between raters excluded from error term

generalizability analyses of these stations, comparing one and two raters per station at a variety of test lengths.

Use of multiple raters per station has only a marginal effect on reproducibility of scores, particularly for longer tests. If enough stations are used to obtain reproducible scores, a sufficient number of raters is automatically sampled. Thus, from a psychometric perspective, a single rater per station suffices[6]. If large numbers of raters are available, it is much better psychometrically to increase the number of stations. For example, in Table 2, a two-hour test with

[6]It may still be desirable to use multiple raters per station to obtain an index of inter-rater agreement and to aid in quality control. This is most easily accomplished through use of "floating raters" who rotate from station to station, usually in the opposite direction of examinees. This approach results in two raters per SP-examinee encounter for a small percentage of encounters across stations.

Table 2: Reproducibility of scores as a function of testing time and number of raters used.[1]

Test Length in Hours[2]	One Rater Per Station	Two Raters Per Station
1	0.43	0.47
2	0.60	0.64
4	0.75	0.78
6	0.82	0.84
8	0.86	0.88
16	0.92	0.93

[1]From Swanson & Norcini, in press
[2]Ten stations per hour

two raters per station requires the same amount of rater time as a four hour test with one rater per station. The former has a reproducibility of 0.64, while the latter has a reproducibility of 0.75, a very significant increase in precision.

Who should rate examinee performance?

Some previous research has suggested that physician-observers are naturally either stringent (hawks) or lenient (doves), and these tendencies are resistant to training (Newble et al., 1980; Ludbrook & Marshall, 1971). In a recent study with checklists of the type used in the University of Limburg Skills Test, Van der Vleuten et al. (in press) studied the impact of training on different types of raters. Trained and untrained groups of non-physicians, medical students and physician faculty rated the videotaped performance of examinees with two SPs. Results indicated that the need for and effectiveness of training varied across groups: it was least needed and least effective for the physician-raters, more needed and effective for the medical students, and most needed and effective for the non-physicians. However, differences in accuracy between groups were nearly eliminated by rater training. Thus, it appears that SPs and other non-physician observers can provide accurate ratings, as long as training is provided.

Inspection of Table 1 supports the results of the Van der Vleuten et al. (in press) study. There are no consistent trends in level of inter-rater agreement as a function of rater characteristics: adequate inter-rater agreement can be achieved through use of SPs or physicians as raters. Intuitively, it does seem likely that SPs and physicians may differ in the aspects of examinee performance that they can rate accurately. Physicians should be more attuned to logical sequencing of questions in history-taking and technical adequacy of some physical examination maneuvers. SPs may be more sensitive to some communication skills (e.g., establishing rapport, sensitivity to patient needs, avoidance of jargon) and better judges of certain examination maneuvers, where feeling what is done provides important information (e.g., palpation generally; pelvic, rectal, and joint examinations specifically).

Practical and educational considerations may be the most important factors in selection of raters. If physicians are available, it may well be desirable for them to serve as raters, because observation of examinees provides useful feedback concerning instructional effectiveness and what trainees can actually do. If physicians are unavailable, SPs provide a logical, less expensive, and, apparently, satisfactory alternative.

Should Checklists or Rating Scales Be Used?
Review of Table 1 indicates that inter-rater agreement is generally better for checklists than rating scales, though agreement is sufficiently good for both that either can be used, particularly in long tests involving large numbers of stations and raters. Presumably, inter-rater agreement is better for checklists, because items are more concretely stated and can be judged more objectively. From the perspective of inter-rater agreement, when checklists and ratings scales are both viable alternatives, checklists are to be preferred. From an educational perspective, checklists also provide better definition of expectations for examinees and more specific feedback on performance. However, a number of areas (e.g., attitudes, aspects of communication skills) are very difficult to assess with checklists without trivializing the aspect of examinee performance to be judged, and such validity considerations are more important then inter-rater agreement. In such situations, use of behaviourally-anchored rating scales is generally advised, though we could find no research basis for this recommendation in the SP literature.

SP-Related Sources of Measurement Error

What are the measurement consequences of using several SPs to play the same role?
Several research groups have investigated this question. Using a subset of the stations from SIU Data Set 1, Dawson-Saunders et al. (1987) compared the performance of groups of examinees who saw different SPs playing the same role. Statistically significant group differences were obtained on SP-based scores for five of the seven cases studied, suggesting that the SP seen had a major influence on scores. The investigators recommended several measures to reduce this influence: using a single SP per case; phrasing patient checklists in lay language; and increasing the training provided for completion of checklists and rating forms. In small scale studies, Hiemstra et al. (1987) and Vu et al. (1987) also found some evidence of differences between SPs trained to play the same role, but, because of small sample sizes, power to detect large differences was limited. However, Cohen et al. (1988), using the University of Toronto Data Set, reported no significant differences between SPs playing the same role, though inter-rater differences were obtained.

The real question is not, however, whether scores resulting from SPs playing the same role differ; impact of using multiple SPs on overall reproducibility is of primary importance. To investigate this issue, Swanson & Norcini (in press), using UMass Data Set 3, conducted analyses of a subset of ten stations where two SPs played the same role. Table 3 compares the reproducibility of tests using one versus two SPs to play the same role as a function of test length.

While scores derived from different SPs playing the same role were often significantly different, total test score reproducibility was not markedly affected by their use. These results assume random assignment of examinees to SPs playing the same role, so that SP-related differences can average out across the test as a whole. This may be approximately true for many SP-based tests given at a single point in time at the same institution, but not for tests given at different times or at multiple sites, as discussed next.

Table 3: Reproducibility of scores as a function of testing time and number of SPs used.[1]

	Data Gathering Skills		Communication Skills	
Test Length in Hours[2]	Same SP	Different SPs	Same SP	Different SPs
1	0.34	0.33	0.59	0.56
2	0.51	0.50	0.74	0.71
4	0.67	0.67	0.85	0.83
6	0.76	0.75	0.90	0.88
8	0.81	0.80	0.92	0.91
16	0.89	0.89	0.96	0.96

[1]From Swanson & Norcini, in press
[2]Three stations per hour

In a well-designed, large scale study using SIU Data Set 2, Tamblyn et al. (1988) investigated the relative accuracy of SPs playing the same role at the two institutions. This provides a particularly strong challenge to use of multiple SPs, since different individuals trained SPs at each school. Videotapes of 252 SP-examinee encounters from Manitoba and 197 SP-examinee encounters from SIU on fifteen common stations were viewed by trained raters, who recorded the accuracy[7] of presentation on case-specific checklists of critical findings. SIU SPs were found to be somewhat more accurate, probably reflecting the more extensive experience of the SIU trainer and SPs. However, average accuracy exceeded 90% at roughly two-thirds of the stations at both institutions. For the remaining stations, accuracy was generally between 80% and 90%, but several SPs had average accuracy scores between 69% and 75%. The investigators concluded that SPs can introduce both random and systematic error into the measurement process and that station- and SP-related factors influencing accuracy merit additional study. Followup analyses (Tamblyn, personal communication) have indicated that the magnitude of SP-related measurement error is

[7]Accuracy was defined as the number of critical findings presented correctly divided by the number of critical findings which could have been presented, given the actions of the examinee, rescaled as a percentage. Thus, accuracy could vary from zero to one hundred percent, with the latter representing a perfectly accurate presentation.

large and systematic, potentially having an impact on individual station scores and overall pass/fail results.

Additional research in this area is highly desirable. Large scale use of SP-based tests depends, in part, on the ability of test developers to train multiple SPs at different sites to play the same roles accurately. Study of training procedures that facilitate "transportability" of stations is badly needed. It seems likely that some aspects of SP performance (consistency in providing a history; affective elements in communication) are transportable, particularly with use of videotape to document how a case should "feel." Other aspects (simulating abnormal physical findings; locating "real" SPs with comparable abnormal findings) may prove much more resistant to duplication at multiple testing centers.

Station-Related Sources of Measurement Error

In virtually all efforts to assess clinical competence, examinee performance on one case is a poor predictor of performance on other cases. This has been termed the content-specificity of clinical competence (Elstein et al., 1978). This phenomenon has been observed across measurement techniques, including written and computer-based simulations (Swanson et al., 1987; Norcini & Swanson, in press), vignette-based short answer tests (De Graaff et al., 1987), chart audits (Erviti et al., 1980), oral exams (Swanson, 1987), and SP-based tests (Van der Vleuten et al., 1988). Long tests, including large samples of cases, are necessary as a consequence, and careful determination of the number of SP-based stations required to obtain reproducible results is clearly merited.

How much testing time is required to obtain reproducible scores?

The studies reviewed in this paper varied substantially in the methods used to estimate and report results of reliability analyses. To achieve greater comparability across studies, we have reanalyzed reported results using generalizability theory to estimate reproducibility of scores within a consistent framework. Because total testing time and time per station varied across studies, our analysis focussed on reproducibility as a function of *testing time*, rather than *number of stations*.

Table 4 presents the results of the analysis. Generalizability coefficients in the table are analogous to coefficient alpha; they are best thought of as the expected correlation between scores derived from similar, but not identical exams using a different sample of stations of the indicated size. Depending upon the purpose of testing, a value of 0.80 is generally viewed as the minimum acceptable level of reproducibility. Coefficients in Table 4 are indicators of overall reproducibility: they reflect all sources of measurement error operating in the testing situation. Because of differences in study and test design, estimation procedures, and results reported, table entries should be viewed as approximate and only roughly comparable.

Although results vary from study to study, reflecting the diversity of skills and examinees assessed, some trends are evident. First and foremost, content specificity is a serious problem. A minimum of three to four hours of testing time is necessary to obtain even minimally reproducible scores. This finding is almost invariant across the range of station formats and skills assessed. The testing time actually used was only adequate in one study (University of

Table 4: Reproducibility of scores as a function of testing time.

Data Set	Test Length in Hours							Average Station Length in Minutes
	1	2	3	4	6	8	12	
Adelaide	0.43[1]	0.60	0.69	0.75	0.82	0.86	0.90	6
CFPC	0.53[1]	0.67	0.77	0.82	0.87	0.90	0.93	15
ECFMG	0.51[1]	0.68	0.76	0.81	0.86	0.89	0.93	20
Limburg	0.54	0.69[1]	0.77	0.82	0.87	0.90	0.93	15
UMass Data Set 1	0.59	0.74	0.81[1]	0.85	0.90	0.92	0.94	30
UMass Data Set 2	0.50	0.67[1]	0.75	0.80	0.86	0.89	0.92	10
UMass Data Set 3	0.34	0.51	0.61	0.67[1]	0.76	0.80	0.86	15
NBME	0.53	0.69	0.77	0.82[1]	0.87	0.90	0.93	20
SIU Data Set 1	0.19	0.31	0.41	0.48	0.58	0.68	0.73[1]	40
UTMB Data Set 1	0.24	0.38	0.49[1]	0.56	0.66	0.72	0.79	10
UTMB Data Set 2	0.19	0.32	0.41[1]	0.48	0.58	0.65	0.74	10
UTMB Data Set 3	0.31	0.47[1]	0.57	0.64	0.73	0.78	0.84	10
Toronto	0.65	0.79	0.85[1]	0.88	0.92	0.94	0.96	6

[1]Approximate testing time used (rounded to nearest whole hour)

Toronto). Though adequate for research purposes and psychometric analysis, testing time was too short to yield reproducible SP-based scores in the remaining studies.

Second, generalizability coefficients are lowest for those studies in which written followup components were included in scores and testing time (SIU and UTMB data sets). In part, this reflects the extra (doubling of) testing time per station which the written components required. However, even if testing time per station were halved, the coefficients would remain relatively low. This indicates that linking written followup questions to SPs reduces generalizability. In part, this is a natural consequence of a shift in what is measured and in the meaning of the scores. The SP component of a station generally provides measurement information concerning data-gathering and communication skills. Follow-up written questions, on the other hand, usually focus on interpretation of findings, differential diagnosis, laboratory utilization, and treatment. Purely

SP-based scores are more homogeneous measures of specific skills. Inclusion of written components broadens the meaning of scores, but at substantial cost in terms of testing time requirements.

Linking SP-based and written components results in very inefficient sampling of content for the skills measured by the latter. In studies in which the reliabilities of sub-skills were reported (Stillman et al., 1986a, 1986b, 1987; Petrusa et al., 1986; Newble & Swanson, 1988), scores for history-taking, physical examination, and communication skills were consistently more reproducible than scores for differential diagnosis, laboratory utilization, and treatment. Many investigators (including the authors) believe that SP-based testing should be used exclusively for assessment of hands-on clinical skills with patients, with traditional paper-and-pencil tests used to assess other components of competence (Stillman et al., 1986a, 1986b, 1987; Swanson, 1987; Van der Vleuten et al., 1989).

How long should individual stations be -- which formats use testing time most efficiently?

Aside from lower reproducibility for stations including written components, there is no obvious relationship in Table 4 between testing time per station and generalizability of scores. Both short and long stations are effective (and ineffective), depending upon the study. Apparently, longer stations tend to yield more measurement information, but the fact that more short stations can be completed in a fixed amount of time completely compensates. There may well be boundaries to this compensatory effect. Tests using very long stations (e.g., more than one hour each) would probably yield less reproducible scores, even controlling for total testing time, because other sources of measurement error (i.e., raters and SPs) would be sampled less extensively. Very short stations (e.g., one or two minutes each), might make efficient use of testing time, but the clinical tasks which can be performed in such a brief time period may not be particularly interesting from an assessment and educational perspective.

In general, selection of station length and format should be considered from the perspective of *content validity*, not reproducibility. For example, a test developer might be interested in the ability of a junior student to perform a complete history and physical examination on a healthy patient; the required format is implied, and a one-hour station length might be appropriate. The first step in development of SP-based tests should be identification of the skills to be assessed. The tasks to be used at individual stations follow naturally, and these constrain decisions regarding station format and length. When SP-based testers choose sharply different station formats and lengths, they are usually interested in assessing different skills.

Reproducibility of Domain-Referenced Test Scores

Most studies to date have adopted (often implicitly in selection of reliability estimation procedures) a norm-referenced framework for score interpretation. That is, scores are given meaning by reference to the performance of other examinees (e.g., an examinee's score is one standard deviation below the mean, in the 95th percentile, etc.), and reproducibility is high to the extent that tests differentiate examinees and allow fairly precise rank-ordering. For SP-based

tests, it seems natural and desirable to use a domain-referenced framework for score interpretation, where an examinee's score is interpretable in absolute terms (e.g., an examinee's score indicates that he/she can take an adequate history for 80% of common ambulatory problems or can perform 90% of the items on a list of physical examination maneuvers).

What happens to reproducibility if a domain-referenced testing perspective is adopted?
Table 5 provides a comparison of reproducibility for norm- versus domain--referenced score interpretation at various test lengths for those studies where the necessary statistical information (variance components) was reported. Reproducibility for domain-referenced score interpretation is lower than for norm-referenced, because differences in test form difficulty affect the former, but not the latter: easier/harder tests affect the accuracy of absolute scores, but not the rank-ordering of examinees. The drop in reproducibility is substantial, and longer tests are necessary as a consequence. Differences in reproducibility do tend to decrease, however, as test length increases, since test forms tend to become more similar in difficulty as the number of stations increases.

Table 5: Reproducibility of scores as a function of test length and score interpretation.

Data Set	Score Interpretation	Test Length in Hours						
		1	2	3	4	6	8	12
Adelaide	Norm-referenced	0.43[1]	0.60	0.69	0.75	0.82	0.86	0.90
	Domain-referenced	0.34	0.51	0.61	0.67	0.76	0.81	0.88
Limburg	Norm-referenced	0.54	0.69[1]	0.77	0.82	0.87	0.90	0.93
	Domain-referenced	0.42	0.59	0.68	0.74	0.81	0.85	0.90
UMass Data Set 3	Norm-referenced	0.34	0.51	0.61	0.67[1]	0.76	0.80	0.86
	Domain-referenced	0.21	0.34	0.44	0.51	0.61	0.68	0.76
SIU Data Set 1	Norm-referenced	0.19	0.31	0.41	0.48	0.58	0.65	0.73[1]
	Domain-referenced	0.10	0.19	0.26	0.32	0.41	0.48	0.58
Toronto[2]	Norm-referenced	0.54	0.70	0.78	0.83	0.88[3]	0.91	0.93
	Domain-referenced	0.41	0.58	0.68	0.74	0.81	0.85	0.89

[1]Approximate testing time used (rounded to nearest whole hour)
[2]Written followup here included, because the necessary variance components for domain-referenced estimation were only reported for the combined data-set
[3]Actual testing time was 5 hours

Reproducibility of Pass/Fail Decisions

The discussion of reproducibility and testing time requirements has, thus far, focussed on the *reproducibility of scores*. For many applications of SP-based

tests, reproducibility of scores is not really important: *reproducibility of pass/fail decisions* is crucial. For example, SP-based tests given after completion of required clerkships and before medical school graduation often focus on history-taking and physical examination skills in an effort to ensure that these skills have been mastered to the level required for postgraduate training (Newble et al., 1978; Stillman and Swanson, 1987; Stillman et al., 1987; Williams et al., 1987). In such testing situations, while it is obviously desirable to estimate an examinee's ability precisely, the basic issue is whether ability exceeds the mastery point, and the reproducibility of pass/fail decisions is of primary importance both practically and psychometrically.

What happens to reproducibility if a mastery-testing approach is adopted?
If most examinees perform well relative to the pass/fail point, fairly short tests can still yield reproducible pass/fail decisions, particularly for examinees at upper ability levels. Swanson & Norcini (in press) and Colliver et al. (1989) provide concrete illustrations. In such situations, use of "sequential testing" procedures may be particularly advantageous. In this approach to assessment, a brief screening test is given initially to all examinees. Those who perform well relative to the pass/fail point are excused from further testing with a passing result. The test is continued for the remaining examinees, concentrating testing time and resources on the "close call" decisions in the vicinity of the passing score. Swanson and Norcini (in press) provides a hypothetical example based upon the University of Adelaide Data Set.

What methods have been used to set standards for SP-based tests?
Implicit in the mastery testing approach is the use of absolute standards in making pass/fail decisions. Unfortunately, SP-based researchers have done almost no work on development of absolute standard setting procedures analogous to those used for written exams (Livingston & Zieky, 1982). Most researchers have not had to confront the standard setting problem, because test scores either did not count or were combined with other assessments. When pass/fail standards were needed, typically a relative standard was set (e.g., two standard deviations below the mean), the score distribution was inspected, looking for "gaps," or a pass/fail point was selected arbitrarily. Given the current interest in use of SP-based tests as graduation exams or as a component of licensing procedures, development of better standard setting procedures should have a high priority.

Equating Scores on Alternate Forms of SP-Based Tests
In many situations where SP-based tests are used, several equivalent forms of a test are developed, similar in overall content and format, but with different stations and/or SPs on each form. Most commonly, multiple forms are required for security reasons when testing is spread out over time, either because the

number of examinees is large[8] or because several cohorts (e.g., successive clerkships or graduating classes) are to be tested. Regardless of the reason, whenever multiple test forms are used, they are unlikely to be equivalent in level and range of difficulty, and any direct comparison of scores would be unfair to those examinees tested with more difficult forms (Angoff, 1984). A variety of statistical procedures, termed *equating methods*, are used with written tests to deal with this problem (Petersen et al., 1989). Relatively little work has been done to adapt these procedures for SP-based assessment, despite the importance of equating for large scale testing applications.

What procedures have been used to equate scores on SP-based tests?
For the most part, users of SP-based tests do not equate scores on alternate test forms. While multiple forms are commonly used (often with the same stations but different SPs), scores are not adjusted for differences in form difficulty. The problem is simply ignored: scores are interpreted as if they are on the same scale. However, a few investigators have developed some "rough and ready" procedures for coping with differences in form difficulty.

At the University of Adelaide, two forms of the Clinical Test are used each year. Examinees are randomly assigned to forms, which should result in roughly equivalent ability groups taking each form. After test administration, the mean score on each form is calculated, and the difference between them is added to the score of each examinee taking the more difficult form (Newble, 1989). When used with written tests, this procedure is termed *mean equating* (Kolen, 1988). Similar stations are also developed in pairs and randomly assigned to the two forms; this procedure should result in forms that are similar, though not identical, in difficulty.

In SP-based tests given at the University of Massachusetts (Data Sets 2 and 3), for security reasons separate test forms are typically constructed from a common pool of stations according to a fixed blueprint. After test administration, scores on each station are standardized to a mean of 500 and a standard deviation of 100, averaged across stations for each examinee, and restandardized across examinees. This procedure also results in a type of mean equating that adjusts for differences in form difficulty, assuming examinees and stations are randomly assigned to test forms (not quite true for UMass test administrations).

Other researchers (e.g., Cohen et al., 1988; Petrusa et al., 1987a, 1988; Williams et al., 1987; Stillman et al., 1987) have investigated trends in scores obtained on different dates of test administration, generally to determine the impact of possible breaches of security. The same analyses can be interpreted in terms of differences in form difficulty. In general, small, non-systematic differences have been observed. However, problems in test security, the ability

[8]Some institutions (e.g., University of Limburg) train several SPs for each station role and create several "replications" of the same test form, using the same stations but different SPs and raters in each replication. This allows a larger number of examinees to be tested concurrently. However, because raters and SPs can differ, despite use of identical station content in each replication, this approach to test administration should still be viewed as involving multiple test forms.

of examinee groups, and the difficulty of test forms are all confounded in these analyses, so results are impossible to interpret.

What other procedures might be used to equate scores on SP-based tests?
A variety of equating procedures have been developed for written tests (Petersen et al., 1989). Several of these could be adapted for use in SP-based assessment, at least for situations where large numbers of examinees are tested with each form. The various forms of common-items linear equating would probably be the simplest and most practical to use. Using this approach, different test forms would include some common stations. Relative performance on common stations by examinee groups taking different test forms provides a basis for estimating group ability and heterogeneity, independent of form difficulty and discrimination. These estimates are then used as a basis for adjustment of scores on alternate forms (Angoff, 1984). Other equating procedures (equipercentile equating; methods based upon use of item response theory) might also be applied; in general, the sample size requirements for these procedures (at least several hundred examinees per form) are too large, however.

If assessment takes place within a mastery-testing framework, it may prove more practical to *identify equivalent pass/fail points* on alternate test forms, rather than attempting to *equate scores*. If primary focus is on pass/fail decisions, this approach could permit appropriate adjustment for differences in test forms without requiring large numbers of examinees. However, it will be necessary to develop improved standard setting procedures before investigating this alternative.

Validity of SP-Based Test Scores

Validity refers to "the accuracy of a prediction or inference made from a test score" (Cronbach, 1971). It is not a property of the test itself, but of interpretations based upon test scores. Thus, the same test can have many validities, depending upon how it is used and how scores are interpreted. Making matters worse, there is typically no "gold standard" with which scores can be compared; this is surely true for SP-based tests. As a consequence, validation requires the accumulation of evidence across a series of studies.

Traditionally, validation studies take one of three forms: 1) study of differences in group performance (e.g., comparison of scores received by examinees at different point in their training); 2) study of the relationships between scores and other measures (e.g., correlations with written test scores and ratings of clinical performance); and 3) study of test content (test blueprint, clinical tasks posed to examinees, items included on checklists, etc.). After a brief discussion of procedures for scoring stations, the following subsections review work in each category. The last subsection outlines some validation studies that we believe are needed.

Procedures for Scoring SP-Based Tests

Since the essence of validity is the accuracy of inferences based upon test scores, it seems appropriate to begin discussion of validity by commenting on procedures used for scoring SP-based stations and tests. Interestingly, published articles rarely comment on scoring procedures, beyond stating that scores were calculated as the percentage of possible points obtained on checklists and/or rating scales. Often, it is unclear how checklists were developed, how individual items were weighted in calculation of station (sub)scores, and how station (sub)scores were aggregated to obtain composite scores. Thus, material in the remainder of this section is based predominantly on speculation, rather than results of research.

What items should station checklists and rating forms include?

Inspection of data-gathering checklists used by different groups reveals remarkable diversity. Checklist length varies from a few items to several dozen or more. On checklists used for history-taking, some groups list questions that examinees should ask; others list answers that SPs provide. Intuitively, the latter seem easier to use, since roughly equivalent questions may be asked in several ways at varying levels of specificity, but the information provided is relatively unambiguous. Checklists used for physical examination generally list examination maneuvers, though the level of specificity varies considerably (e.g., from "examines the abdomen" to "palpates the right upper quadrant for the liver" to a list of several discrete steps for each quadrant). Similarly, rating forms for communication skills vary from a single global item to several dozen items concerning discrete skills. Individual items may be fairly concrete ("maintains good eye contact"), fairly abstract ("establishes good rapport"), or related to behavioral intentions ("I would recommend this examinee to a friend). In part, differences in checklists and rating forms may reflect differences in the focus of assessment (more detailed lists for interpersonal skills and physical diagnosis courses; less detailed lists for exams required for graduation from school). Reproducibility of test scores appears to be fairly invariant across the various rating form and checklist formats; validity of scores may not be. Some systematic research on the content and format of checklists and rating forms seems highly desirable.

How should items and subscores be weighted in calculation of station scores?

Relatively little research on this question has been reported in the standardized patient literature. Stillman et al. (1986a) explored the use of weighting in calculation of checklist scores and found little impact on reproducibility or validity. These findings parallel those obtained in studies of written patient management problems (Swanson et al., 1987) and in psychometric studies of weighted composites more generally. Given a particular checklist or rating form, as long as items and subscores are positively intercorrelated, weighting is unlikely to have much impact (Dawes and Corrigan, 1974). This does *not* mean that all scoring approaches will yield similar results, however. It simply indicates that research should concentrate on the items to be included on checklists and rating forms in the first place, rather than on alternative weighting systems.

What scores and subscores should be reported to examinees?
The answer to this question is complex: it depends upon the purpose of testing, the reproducibility and validity of scores, and security considerations. If the purpose of testing is primarily formative -- to identify strengths and weaknesses of examinees -- providing extensive feedback on performance may be desirable. Examinees taking written tests are often provided with the test materials after scoring is complete so they can study material related to items they answered incorrectly. In SP-based tests, giving examinees copies of their checklists and rating forms both provides feedback on their performance and defines performance criteria used in grading. Students may then "study to the test," which can be desirable and effective, if mastery of specific skills included in the test form is desired. However, if the resulting performance gains are specific to a particular test form, rather than generalized, this kind of feedback is less useful and poses security problems for reuse of stations.

When the purpose of a test is primarily summative -- to make grading or pass/fail decisions -- providing extensive feedback may be counterproductive. From a security standpoint, reuse of stations becomes problematic. Further, in most situations, only the total test score is sufficiently reproducible to be meaningful. Reporting a profile of subscores is particularly ill-advised, since individual subscores in the profile are likely to be very unstable.

Differences in Group Performance

Do examinees at different points in training perform differently?
There have been relatively few studies of the performance of different groups on SP-based tests. In UMass Data Set 1, the performance of internal medicine residents improved as they progressed through training: third year residents performed better than second year residents, who performed better than first year residents (Stillman et al., 1986a, 1986b). In the same study, residents from training programs with stronger reputations performed significantly better than residents from less prestigious programs. In UTMB Data Set 2, first and second year residents in internal medicine performed significantly better than junior medical students (Petrusa et al., 1986). Other small scale studies have observed similar trends (Newble et al., 1981; Stillman et al., 1982; Robb and Rothman, 1985). Thus, the results of studies of differential group performance provides some support for the validity of SP-based test scores.

This evidence, in isolation, is hardly compelling. Any well-constructed twenty-minute multiple choice test differentiates groups at different levels of training, simply on the basis of differences in knowledge. Further, it is usually unclear *how much* groups should differ, so such studies can only yield very imprecise information. Results are only conclusive when they are negative; when expected differences are not obtained, strong evidence of *invalidity* is provided. More sophisticated studies are needed to provide supportive evidence.

Relationships with Other Measures of Clinical Competence

Table 6 summarizes the correlations between SP-based test scores and a variety of other indices of clinical competence, as reported in the studies included in the review. The middle column provides observed correlations. These vary non-

systematically from study to study, in part because they are "attenuated" (reduced in magnitude) by measurement error. For short, unreliable tests, a sizable reduction can occur (e.g., the rows for University of Adelaide). The rightmost column provides "true" (disattenuated) correlations. The effect of measurement error has been eliminated from these correlations statistically; they can be viewed as the correlations that would be obtained if the tests were very long (perfectly reliable). Consequently, they provide a better index of the true strength of relationship between the measures involved and should be easier to interpret.

What is the relationship between SP-based test scores and other measures?
True correlations between scores on SP-based and multiple choice tests vary extensively, from near zero to one, though the average value is fairly high. True correlations between SP-based scores and ratings of clinical performance are also moderately high. These results are not particularly surprising. The performance of better trainees should exceed those of poorer trainees across a wide range of content and skills, and there is a common core of clinical knowledge that underlies performance on most tests, regardless of the skills measured. Given similar educational goals and opportunities to learn, better (brighter, more highly motivated, self-directed) students will outperform poorer students, and, as time in training passes, this effect should increase in size (e.g. Van der Vleuten et al., 1989). This should be true, *regardless of the format of the achievement test used,* as long as a reasonably broad range of content and skills is covered. In factor analytic studies of the structure of clinical competence, this results in models including only a single general factor (Maatsch, 1980, 1987; Maatsch & Huang, 1986).

Some might view the moderately high true correlations between scores on SP-based tests and traditional assessments as an indication that the same trait is being measured. This is simply wrong. High correlations indicate that tests rank-order examinees similarly, without saying anything about the specific skills measured (Swanson, 1987). Within a norm-referenced framework and from a purely psychometric perspective, high correlations do indicate that tests can be used interchangeably without affecting which examinees pass and fail. This is reassuring information in a sense; most examinations used for grading, licensure and certification are written, norm-referenced tests, not direct assessments of clinical skills. Within domain-referenced and mastery testing frameworks, the equivalence of SP-based and traditional tests disappears, because the absolute level of performance is of interest, not just the relative ranking of examinees (Newble & Swanson, 1988).

Thus, for the most part, studies of the relationship between SP-based tests and other measures provide supportive evidence for the validity of test scores, though almost any results could be interpreted positively. (For example, in UMass Data Set 1 and 2 where lower true correlations were observed, the investigators concluded that SP-based tests measure important aspects of clinical competence not tapped by traditional measures.) This highlights a weakness of broad, unfocussed correlational studies of validity: results can be interpreted in a variety of ways, with almost any findings viewed as positive or negative.

Table 6: Relationship between SP-based scores and other measures.

Data Set/Measure	Observed Correlation	True Correlation[1]
Adelaide		
Multiple Choice Test (Locally developed)	0.33	0.68
Non-SP Skills Test	0.35	1.00
Written Followup (Short Answer Test)	0.40	0.88
Limburg		
Multiple Choice Test (Locally developed)[2]	0.63	0.77
Written Followup (Multiple Choice Test)[2,3]	0.62	0.77
UMass Data Set 1		
Multiple Choice Test (ABIM Certifying Exam)	0.24	0.29[4]
Clinical Ratings	Non-Significant	-
Months of Residency Training	0.32	0.37[4]
Self-Ratings	Non-Significant	-
UMass Data Set 2		
Multiple Choice Test (NBME Part I)	0.19	0.24[4]
Multiple Choice Test (NBME Part II)	0.27	0.34[4]
Clinical Ratings	0.44	0.50[4,5]
Written Followup (Short Answer & Multiple Choice Test)	0.26	0.36[4]
UMass Data Set 3		
Multiple Choice Test (NBME Part I)	0.10	0.13[4]
Multiple Choice Test (NBME Part II)	0.22	0.28[4]
Clinical Ratings	0.25	0.31[4,5]
Self-Ratings	0.08	0.10[4,5]
Written-followup (Pattern Recognition Test)	0.22	0.32[4]
NBME		
Multiple Choice Test (NBME Behavioral Science Subtest)	0.37	0.46[4]
Multiple Choice Test (NBME Psychiatry Subtest)	0.31	0.40[4]
Multiple Choice Test (Other NBME)	Non-Significant	
SIU Data Set 1		
Multiple Choice Test (NBME Part I)	0.53	0.63[4]
Multiple Choice Test (NBME Part II)	0.51	0.60[4]
Clinical Ratings	0.65	0.75[4,5]
SIU Data Set 2		
Multiple Choice Test (NBME Part II)	0.63	0.82[4]
Clinical Ratings	0.52	0.66[4,5]
UTMB Data Set 1		
Multiple Choice Test (NBME Medicine Subtest)	0.43	0.64
Clinical Ratings	0.46	0.73

Table 6 continued

UTMB Data Set 2		
Multiple Choice Test (ABIM Certifying Exam)	0.24	0.35
Clinical Ratings	0.00	-
Months of Residency Training	0.31	0.49
UTMB Data Set 3		
Multiple Choice Test (NBME Medicine Subtest)	0.64	1.00
Clinical Ratings	0.37	0.56[5]
Toronto		
Multiple Choice Test (Locally Developed)	0.43	0.50
Written followup (Short Answer Test)	0.69	0.91

[1]Entries corrected for measurement error (disattenuated) in both scores, unless otherwise noted
[2]Reported in Van der Vleuten et al. (1989)
[3]Administered two weeks after SP-test
[4]No values were reported; entries were approximated from available results
[5]Estimate is disattenuated for unreliability in SP-score only

This problem, among others, led Ebel (1961) to state:

> *"Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few."* (p. 640)

Content Validity of SP-Based Tests

Ebel (1965) suggested that validation procedures can be divided into two categories: direct (primary) and derived (secondary). The latter term applies to the studies reported in the previous subsection, where the focus was on the relationship between test scores and other measures. In contrast, direct validation procedures investigate the extent to which the tasks posed by test items faithfully represent the real-world tasks of measurement interest. These validation procedures depend upon rational analysis of the test in relation to the domain of interest, rather than empirical, statistical evidence. For achievement tests (as opposed to personality or aptitude tests), Ebel clearly believed that direct validation is most important. He argued that validity can be "built into" a test through careful operational definition of the tasks and content to be measured. Kane (1982) and Frederiksen (1984) expressed a similar point of view: content validity should be the major concern in achievement testing.

Are SP-based tests content valid?

On the surface, it appears that SP-based test scores should be content valid because the "testing tasks" posed resemble real-world clinical tasks. However, it is unclear whether SP-based tests are actually constructed as systematically as Ebel would demand. To determine this, it would be necessary to review the

domain definitions developed by test authors, determine whether those domains have been representatively sampled, study the scoring methods, and conduct generalizability studies (since content validity depends upon having a large enough sample of items). Test and station construction procedures are probably less systematic than this; this was certainly true in most of the studies in which we participated directly. Careful description of the domain(s) to be tested and systematic development of sampling plans and test blueprints are needed. Content validity follows from attention to these details.

Needed Validation Studies

SP-based tests seem particularly amenable to direct validation. They consist of a series of high fidelity simulations, so there should be little concern about differences in the tasks posed by individual "items" and those posed by the real world[9]. If the skill/content domain to be tested is carefully defined and sampled, validity should be built in through the test construction process, as long as a sufficiently large sample of SP cases is included. However, several "threats to validity" remain to be studied.

First, it is unclear if scoring procedures accurately translate examinee behaviour into appropriate, meaningful scores. In general, published reports vaguely describe scoring as calculation of "percentage of possible points" obtained. The validity of such scores depends upon the appropriateness of the items, the weighting attached to each, and other factors. The potential for omitting important items and including unimportant ones is great. The former penalizes examinees who take indicated actions that are not listed; the latter rewards examinees who are unjustifiably thorough, a common problem in scoring written and computer-based patient management problems (Swanson et al., 1987). Research investigating commonly used checklist formats and scoring procedures is badly needed. One obvious approach is to use several methods to develop checklists and ratings forms for scoring the same stations and compare the results. Complementary research would check the correspondence between station scores and ratings of examinee behaviour provided by expert observers who are unfamiliar with scoring algorithms. Multiple observers could be used, with agreement between them providing an index of the agreement one would hope to see with scoring procedures. If done on a broad range of stations, this might be termed the *microscopic approach* to test validation. Given that the scores on individual stations are valid, total test scores should also be valid, as long as stations are representatively sampled from the defined domain, and the test is sufficiently long to yield reproducible scores.

Second, examinee scores may well be affected by a mismatch between their perceptions of the tasks posed by stations and those intended by test develop-

[9]Several investigations directly compared performance of SPs and real patients. These have uniformly indicated that SP-behavior is comparable to real patients (Norman et al., 1982; Norman et al., 1985; Sanson-Fisher & Poole, 1980; Owen & Winkler, 1974; Burri et al., 1976; Renaud et al., 1980; Rethans & Van Boven, 1987).

ers[10]. (This might be termed the "guess what I want you to do" problem.) The impact of a mismatch may well be most serious when short stations are used, since examinees must quickly determine what to do, without time to readjust if they guess wrong. Interview studies with examinees should shed some light on this potential problem: it is always useful to ask examinees why they behaved as they did.

A related problem follows from the time pressure under which SP-based tests are commonly administered. Developers of written tests try to obtain information on the time required for examinees to comfortably complete the test. If insufficient time is allowed (i.e., the test is "speeded"), score interpretation is more difficult. A straightforward method for investigating "speededness" of SP-based tests would involve comparing performance on the same stations under varying time conditions. Given the diversity of stations and station formats, it would not be surprising if this factor were important in examinee performance, and more specific guidelines for station construction could result.

Because station scores are dependent upon the judgment of observers, characteristics of examinees unrelated to clinical skills could influence those judgments. Such characteristics include, age, gender, ethnicity, accent, and appearance of examinees. This possibility is difficult to investigate because these characteristics can covary with clinical skills in unknown ways. However, it should be possible to train some "standardized examinees" who behave in a consistent fashion, making possible systematic investigation of the influence of such characteristics on rater judgments. Videotapes in which the clinical skills of simulated examinees are systematically varied in relation to gender, race, etc. could also yield useful information. Such "bias studies" may be especially important if SP-based tests are to be used for assessment of the clinical skills of foreign medical graduates.

Educational Impact of SP-Based Tests

What are the educational consequences of SP-based tests?

Several authors have emphasized the importance of considering the impact of assessment methods on education, both in general (Frederiksen, 1984; Entwistle, 1981) and for SP-based tests specifically (Newble & Jaeger, 1983; Stillman & Swanson, 1987; Bouhuijs et al., 1987; Van der Vleuten et al., 1989). Clinical skills instruction is typically provided using a fairly non-systematic, master-apprentice approach. Curricular goals, instructional quality, and educational material (i.e., patients) can vary greatly from school to school, hospital to hospital, and master to master (Stillman & Swanson, 1987). SP-based tests are hypothesized to counteract these problems, both by influencing

[10]Because of this problem, examinees should be provided with extensive information about the purpose and format of the test *in advance of the examination*. "Practice" tests with feedback on performance, written materials including sample checklists and rating forms, and videotapes of stations from previous exams should all be useful, particularly if examinees vary in their familiarity with SP-based testing.

the faculty's teaching and the students' learning activities. The argument runs as follows.

In order to develop an SP-based exam, it is necessary for faculty to reach a consensus on what should be learned. Domain definition and blueprinting, if taken seriously, require specific delineation of the range of clinical situations and skills which trainees should have mastered. Station construction requires concrete definition of performance criteria. Reaching a consensus on these items should eventually lead to more standardized instructional experiences and learning outcomes, particularly if the consensus is carefully communicated to both faculty and students (Bouhuijs et al., 1987). Trainees view assessment methods as indicators of what faculty believe is important, and deliberate manipulation of exams can exert a major influence on the learning activities.

The only empirical work on the educational impact of SP-based tests has been at the University of Adelaide. The original motivation for the development of the Clinical Test (see subsection on University of Adelaide Data Set 1) was a general faculty perception that students were spending a disproportionate amount of time studying for written exams relative to clinical work on wards (Newble, 1986, 1987, 1988; Newble & Jaeger; 1983). The Clinical Test was introduced to improve the congruence between the educational goals of the medical school and the assessment methods used, anticipating that a shift in students' learning activities might result. To investigate the impact on student study habits, questionnaires were sent to students before and after the introduction of the Clinical Test component of the final examination (Newble & Jaeger, 1983). Questionnaire results indicated that the Clinical Test had a dramatic impact on how students spent their time, decreasing efforts to prepare for the Theory Test and increasing ward-based learning activities. In addition, students reported a generally high level of satisfaction with the Clinical Test and rated it as substantially more relevant than the Theory Test to the work of an intern. These results have persisted since the Clinical Test was introduced (Newble, 1988).

Discussion

The purpose of this paper was to review psychometric research on SP-based tests. In this final section, we summarize the major conclusions reached in the review, present some suggestions for improved use of SP-based tests, and provide some methodological observations and recommendations.

Summary of Conclusions

This review was divided into three major areas: reproducibility of scores and pass/fail decisions; validity of score interpretation, and educational impact of tests. This summary also follows that organization.

Reproducibility of SP-Based Scores and Pass/Fail Decisions. Lack of inter-rater agreement in scoring examinee behaviour, inconsistency in SP performance, and variation in examinee performance across stations all affect the

reproducibility of scores. Inter-rater reliability is adequate, regardless of rater background, as long as requisite training is provided. Use of multiple SPs playing the same patient role does not generally reduce reproducibility very much. Variation in examinee performance across stations has the largest impact on the reproducibility of scores. In most testing situations, if a sufficiently large sample of stations is included, the resulting sample of raters and SPs will also be large enough to obtain reproducible scores. It appears that exceptions can occur in large scale testing situations, when examinees are tested at different times and/or at different sites -- when there is major departure from random assignment of raters and SPs to examinees. Four to eight hours of testing time are required to obtain reproducible scores for hands-on clinical skills; longer tests are required if stations include written questions linked to SPs. Otherwise, station format and testing time per station seem unrelated to reproducibility. Domain-referenced score interpretation requires longer tests than norm-referenced. Mastery-testing, in which only pass/fail results are of interest, has the potential to reduce test length requirements and costs, particularly when combined with sequential test administration. However, better standard setting procedures must be developed in order to realize this potential. Similarly, more work is needed on procedures for statistically adjusting (equating) scores obtained on alternate test forms that differ in difficulty and discrimination.

Validity of SP-Based Test Scores. Results of validation studies have, for the most part, been encouraging, though not particularly informative. Groups at different stages of training obtain appropriately different scores, and relationships between SP-based scores and traditional measures of clinical competence are fairly strong. Content validity of SP-based tests should be particularly good, because of the realistic clinical tasks included as stations, though additional attention to domain definition and blueprinting seems merited. Additional efforts should be devoted to test validation; this could include research on scoring procedures, examinee perceptions of tasks posed by stations, effect of station speededness, and rater bias.

Educational Impact of SP-Based Tests. Aside from Newble's efforts at the University of Adelaide, there has been little empirical work on the educational impact of SP-based tests. More research in this area is needed, since the hypothesized educational impact of SP-based tests has been a major factor in their increased use, despite high costs and psychometric shortcomings.

Suggestions for the Improved Use of SP-Based Tests

A number of practical suggestions for improving SP-based tests and conserving testing resources follow directly from the psychometric conclusions emerging from the review. These are outlined below.

1. Do not try to interpret scores on short SP-based tests; they are not sufficiently reproducible. Profiles of subscores based upon a small number of stations or short segments of each station are also unstable and should generally not be reported.

2. There is no need to use more than one rater per station. If extra raters are available, increase the number of stations used.
3. The decision to use non-physicians (usually SPs) or faculty physicians as raters has practical and educational elements. If faculty physicians serve as raters, they receive useful feedback on instructional effectiveness through observation of examinees. SPs can also serve effectively as raters, given adequate training; often, they are more readily available and less expensive.
4. SP-based tests should emphasize assessment of hands-on skills, such as history taking, physical examination, patient education, counseling, etc. Linking the hands-on components to written questions concerning differential diagnosis, laboratory utilization, and treatment should be avoided. If these skills are to be tested, written tests administered separately are preferable.
5. Selection of station formats follows from the hands-on clinical skills to be assessed; reproducibility of scores appears to be almost unrelated to format.
6. Relatively little measurement error is introduced by training multiple SPs to play the same patient role. This approach can reduce training time, increase scheduling flexibility, and allow larger numbers of examinees to be tested concurrently.
7. Procedures for setting pass/fail standards on SP-based tests remain primitive. Do not rely solely on SP-based tests in making major promotion/certification decisions.
8. Use SP-based tests. Despite the early stage of development, the tests measure important skills emphasized in clinical training. Assessment should be congruent with educational goals.

Methodological Observations and Recommendations

Published reports (including our own) were often distressingly vague. Psychometrically significant details of test administration (e.g., whether multiple SPs played the same patient role; how examinees were assigned to raters, SPs, and stations) were often unclear. The format and content of checklists and rating forms and the procedures used for scoring were almost never described. More work is needed in these areas, beginning with better descriptions of the methods already in use.

Procedures used for reliability estimation were sometimes unspecified and, occasionally, appeared to be wrong. Use of generalizability theory in analysis is absolutely required, because multiple sources of measurement error are commonly present. Variance component estimates should always be reported (along with the standard errors of the estimates), so that readers can explore alternative uses of testing resources, and researchers can better integrate results across studies. If multiple subscores are calculated, results of generalizability analyses should generally be reported for all of them.

In validity analyses, both observed and true correlations should be reported, along with reliability information for criterion measures, if available. Observed correlations are often seriously attenuated; without reliability information they are uninterpretable. Correlations between SP-based test scores and traditional measures of clinical competence do not provide particularly useful information. More creative, focussed studies of the validity of SP-based tests are clearly needed.

References

Andrew, B. (1977) The use of behavioral checklists to assess physical examination skills. *Journal of Medical Education, 52*, 589-591.

Angoff, W. (1984) *Scales, Norms, and Equivalent Scores.* Princeton: Educational Testing Service.

Barrows, H., Williams, R. & Moy, R. (1987) A comprehensive performance-based assessment of fourth-year students' clinical skills. *Journal of Medical Education, 62*, 805-809.

Bouhuijs, P., Van der Vleuten, C. & Van Luyk, S. (1987) The OSCE as a part of a systematic skills training approach. *Medical Teacher, 9*, 183-191.

Brennan, R. (1983) *Elements of Generalizability Theory.* Iowa: American College Testing Program.

Burri, A., McCaughan, K. & Barrows, H. (1976) The feasibility of using the simulated patient as a means to evaluate clinical competence of practicing physicians in a community. *Proceedings of the 15th Annual Conference on Research in Medical Education*, 295-299.

Cody, R. (1988) *Additional analysis for the 1987 administration of the clinical skill exam.* Internal Report Educational Commission for Foreign Medical Graduates.

Cohen, R., Rothman, A., Ross, J., MacInnes, A., Domavitch, E., Jamieson, C., Jewitt, M., Keystone, J., Kulesha, D., McCleary, P., Ouchterlony, D., Poldre, P., Robb, K., Rossi, M., Sarin, M., Schwartz, M., Sherman, R. & Shier, M. (1987) Comprehensive assessment of clinical performance. In: Hart, I. & Harden, R. (Eds.) *Further Developments in Assessing Clinical Competence.* Montreal: Heal Publications.

Cohen, R., Rothman, A., Ross, J., Keystone, J., Kulesha, D., MacInnes, A., McLeary, P., Ouchterlony, D., Poldre, P., Robb, K., Rossi, M., Sarin, M. & Schwartz, M. (1988) *A Comprehensive Assessment of Graduates of Foreign Medical Schools.* Internal Report, University of Toronto.

Colliver, J., Verhulst, S., Williams, R. & Norcini, J. (1989) Reliability of performance on standardized patient cases: A comparison of consistency measures based on generalizability theory. *Teaching and Learning in Medicine, 1*, 31-37.

Conn, H. (1986) Assessing the clinical skills of foreign medical graduates. *Journal of Medical Education, 61*, 863-871.

Conn, H. & Cody, R. (Under editorial review) *Clinical Skills Examination of the ECFMG (II).*

Cronbach, L.J. (1971) Test validation. In: Thorndike, R.L. (Ed.) *Educational Measurement.* Washington D.C.: American Council on Education.

Cronbach, L., Gleser, G., Nanda, H. & Rajaratnam, N. (1972) *The Dependability of Behavioral Measurements: Generalizability for Scores and Profiles.* New York: John Wiley and Sons.

Dawes, R. & Corrigan, B. (1974) Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Dawson-Saunders, B., Verhulst, S., Marcy, M. & Steward, D. (1987) Variability in standardized patients and its effect on student performance. In: Hart, I.R. & Harden, R.M. (Eds.) *Further Developments in Assessing Clinical Competence.* Montreal: Can-Heal.

De Graaff, E., Post, G. & Drop, M. (1987) Validation of a new measure of clinical problem-solving. *Medical Education, 21*, 213-218.

Ebel, R. (1961) Must all tests be valid? *American Psychologist, 16*, 640-647.

Ebel, R. (1965) *Measuring Educational Achievement*. Englewood Cliffs: Prentice-Hall, Inc.
Elstein, A., Shulman, L. & Sprafka, S. (1978) *Medical Problem Solving*. Cambridge: Harvard University Press.

Entwistle, N. (1981) *Styles of Learning and Teaching*. Chichester: John Wiley & Sons.

Erviti, V., Templeton, B., Bunce, J. & Burg, F. (1980) The relationships of pediatric resident recording behavior across medical conditions. *Medical Care, 18*, 1020-1031.

Frederiksen, N. (1984) The real test bias: influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Grava-Gubins, I., Rainsberry, P. & Khan, S. (1985a) *A study of the reliability and validity of the formal oral examination*. Internal Report, College of Family Physicians of Canada, March, 1985.

Grava-Gubins, I., Khan, S. & Rainsberry, P. (1985b) *A study of the reliability of the 1985 simulated office orals*. Internal Report, College of Family Physicians of Canada, October, 1985.

Grava-Gubins, I., Khan, S. & Rainsberry, P. (1985c) *A factor analytic study of the 1985 simulated office orals*. Internal Report, College of Family Physicians of Canada, October, 1985.

Grava-Gubins, I., Rainsberry, P. & Khan, S. (1986) *Reliability and factor analytic research on 1986 simulated office oral examinations*. Internal Report, College of Family Physicians of Canada, December, 1986.

Grava-Gubins, I., Khan, S. & Rainsberry, P. (1987) Factor analysis of simulated office oral examinations in family medicine. In: Hart, I.R. & Harden, R.M. (Eds.) *Further Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Grava-Gubins, I., Rainsberry, P. & Kahn, S. (1988) *A study of the structure of the 1987 simulated office oral examinations*. Internal Report, College of Family Physicians of Canada, March, 1988.

Harden, R., Stevenson, M., Downie, W. & Wilson, G. (1975) Assessment of Clinical Competence using objective structured examinations. *British Medical Journal, 1*, 447-451.

Harden, R. & Gleeson, F. (1979) Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 13*, 41-54.

Hart, I., Harden, R. & Walton, H. (Eds.), (1986) *Newer Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Hart, I. & Harden, R. (Eds.), (1987) *Further Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Hiemstra, R., Scherpbier, A. & Roze, B. (1987) Assessing history-taking skills or ... simulated patients' peculiarities. In: Hart, I.R. & Harden, R.M. (Eds.) *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.

Kane, M. (1982) The validity of licensure examinations. *American Psychologist, 37*, 911-918.

Khan, S., Grava-Gubins, I., Rainsberry, P. (1988) *Generalizability analysis of the 1987 simulated office oral examinations.* Internal Report, College of Family Physicians of Canada, March, 1988.

Klass, D., Hazzard, T., Kopelow, M., Tamblyn, R., Barrows, H. & Williams, R. (1987) Portability of a multiple station, performance based assessment of clinical competence. In: Hart, I. & Harden, R. (Eds.) *Further Developments in Assessing Clinical Competence.* Montreal: Heal Publications.

Kolen, M. (1988) Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*, 29-36.

Livingston, S. & Zieky, M. (1982) *Passing Scores.* Princeton: Educational Testing Service.

Ludbrook, J. & Marshall, V.R. (1971) Examiner training for clinical examinations. *British Journal of Medical Education, 5*, 152-155.

Maatsch, J. (1980) *Model for a criterion-referenced medical specialty test.* Final Report Grant No. HS-02038-02, Office of Medical Education Research and Development Michigan State University.

Maatsch, J. & Huang, R. (1986) An evaluation of the construct validity of four alternative theories of clinical competence. *Proceedings of the Twenty-fifth Annual Conference on Research in Medical Education,* Washington, DC.

Maatsch, J. (1987) Theories of clinical competence: The construct validity of objective tests and performance assessments. *Paper presented at the International Conference on Evaluation in Medical Education,* Beer Sheva, Israel.

Neufeld, V., Woodward, C. & Norman, G. (1983) Simulated patients in evaluating of medical education. *Proceedings of the 22nd Annual Conference on Research in Medical Education,* 240-242.

Newble, D. (1986) The assessment of clinical competence - A perspective from "down under". In: Hart, I., Harden, R. & Walton J. (Eds.) *Newer Developments in Assessing Clinical Competence.* Montreal: Heal Publications.

Newble, D. (1986) Improving the clinical and oral examination process. In: Hart, I., Harden, R. & Walton J. (Eds.) *Further Developments in Assessing Clinical Competence.* Montreal: Heal Publications.

Newble, D. (1988) Eight years' experience with a structured clinical examination. *Medical Education, 22*, 200-204.

Newble, D. (1989) Personal Communication.

Newble, D., Elmslie, R. & Baxter, A. (1978) A problem-based criterion-referenced examination of clinical competence. *Journal of Medical Education, 53*, 720-726.

Newble, D., Hoare, J. & Sheldrake, P. (1980) The selection and training for clinical examinations. *Medical Education, 14*, 345-349.

Newble, D., Hoare, J. & Elmslie, R. (1981) The validity and reliability of a new examination of the clinical competence of medical students. *Medical Education, 17*, 165-171.

Newble, D. & Swanson, D. (1988) Psychometric characteristics of the objective structured clinical examination. *Medical Education, 22*, 325-334.

Newble, D. & Jaeger, K. (1983) The effect of assessments and examinations on the learning of medical students. *Medical Education, 17*, 165-171.

Norcini, J. & Swanson, D. (in press) Factors influencing testing time requirements for simulation-based measurements: Do simulations ever yield reliable scores? *Teaching and Learning in Medicine.*

Norman, G., Neufeld, V., Walsh, A., Woodward, C. & McConvey, G. (1985) Measuring physicians' performance by using simulated patients. *Journal of Medical Education, 60*, 925-934.

Norman, G., Tugwell, P. & Feightner, J. (1982) A comparison of resident performance on real and simulated patients. *Journal of Medical Education, 57*, 708-715.

Owen, A. & Winkler, R. (1974) General practitioners and psychosocial problems: an evaluation using pseudopatients. *Medical Journal of Australia, 2*, 393-398.

Petersen, N., Kolen, M., & Hoover, H. (1989) Scaling, norming, and equating. In: Linn, R. (Ed.) *Educational Measurement.* New York: American Council on Education and MacMillan Publishing Company.

Petrusa, E., Guckian, J. & Perkowski, L. (1984) A multiple station objective clinical evaluation. *Proceedings of the Twenty-third Annual Conference on Research in Medical Education*, 211-216.

Petrusa, E., Blackwell, T., Parcel, S. & Saydjari, C. (1986) Psychometric properties of the Objective Clinical Exam as an instrument for final evaluation. In: Hart, I., Harden, R. & Walton J. (Eds.) *Newer Developments in Assessing Clinical Competence.* Montreal: Heal Publications.

Petrusa, E., Blackwell, T., Rogers, L., Saydjari, C., Parcel, S. & Guckian, J. (1987a) An objective measure of clinical performance. *American Journal of Medicine, 83*, 34-42.

Petrusa, E., Blackwell, T. & Ainsworth, M. (1987b) Performance of internal medicine house officers on a short station OSCE. In: Hart, I., Harden, R. & Walton J. (Eds.) *Further Developments in Assessing Clinical Competence.* Montreal: Heal Publications.

Petrusa, E. (1988) *Collaborative Project to Improve the Evaluation of Clinical Competence.* Final Report to the National Fund for Medical Education.

Rainsberry, P., Grava-Gubins, I., Khan, S. (1985) *A factor analytic investigation of the simulated office orals and the formal oral examination.* Internal Report, College of Family Physicians of Canada, April, 1985.

Rainsberry, P., Grava-Gubins, I., & Khan, S. (1987a) Reliability and validity of oral examinations in family medicine. In: Hart, I. & Harden, R. (Eds.) *Further Developments in Assessing Clinical Competence.* Montreal: Heal Publications.

Rainsberry, P., Grava-Gubins, I. & Khan, S. (1987b) *A reliability study of the 1987 simulated office orals in family medicine.* Internal Report, College of Family Physicians of Canada, August, 1987.

Renaud, M. et al. (1980) Practice settings and prescribing profiles: the simulation of tension headaches to general practitioners working in different practice settings in the Montreal area. *American Journal of Public Health, 70*, 1068-1073.

Rethans, J. & Van Boven, C. (1987) Simulated patients in general practice: a different look at the consultation. *British Medical Journal, 294*, 809-812.

Robb, K. & Rothman, A. (1985) The assessment of clinical skills in general internal medicine residents – comparison of the objective structured clinical examination to a conventional oral examination. *Annals of the Royal College of Physicians and Surgeons of Canada, 18*, 235-238.

Sanson-Fisher, R. & Poole, A. (1980) Simulated patients and the assessment of medical students' interpersonal skills. *Medical Education, 14*, 708-715.

Stillman, P., Ruggill, J., Rutala, P. & Sabers, D. (1980) Patient instructors as teachers and evaluators. *Journal of Medical Education, 55*, 186-193.

Stillman, P., Rutala, P., Nicholson, G., Sabers, D. & Stillman, A. (1982) Measurement of clinical competence of residents using patient instructors. *Proceedings of the 21st Annual Conference on Research in Medical Education*, 111-116.

Stillman, P., Swanson, D., Smee, S., Stillman, A., Ebert, T., Emmel, V., and the New England Consortium of Internal Medicine Residency Training Programs (1986a) *Psychometric Characteristics of Standardized Patients for Assessment of Clinical Skills*. Final Report to the American Board of Internal Medicine.

Stillman, P., Swanson, D., Smee, S., Stillman, A., Ebert, T., Emmel, V., Caslowitz, J., Greene, H., Hamolsky, M., Hatem, C., Levenson, D., Levin, R., Levinson, G., Ley, B., Morgan, J., Parrino, T., Robinson, S. & Willms, J. (1986b) Assessing clinical skills of residents with standardized patients. *Annals of Internal Medicine, 105*, 762-771.

Stillman P. & Swanson, D. (1987) Ensuring the clinical competence of medical school graduates through standardized patients. *Archives of Internal Medicine, 147*, 1049-1052.

Stillman, P., Regan, M. & Swanson, D. (1987) A diagnostic fourth year performance assessment. *Archives of Internal Medicine, 147*, 1981-1985.

Stillman, P., Regan, M., Swanson, D., Case, S., McCahan, J., Smith, S., Willms, J., Feinblatt, J. & Finnegan, S. (under editorial review) An assessment of the clinical skills of New England fourth year medical students.

Swanson, D. (1987) A measurement framework for performance-based tests. In Hart, I. & Harden, R. (Eds.) *Further Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Swanson, D., Norcini, J. & Grosso, L. (1987) Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education, 12*, 220-246.

Swanson, D. & Norcini, J. (in press) Factors influencing the reproducibility of tests using standardized patients. *Teaching and Learning in Medicine*.

Tamblyn, R., Schnabl, G., Klass, D., Kopelow, M. & Marcy (1988) How standardized are standardized patients? *Proceedings of the Twenty-seventh Annual Conference on Research in Medical Education*, 148-153.

Templeton, B., Best, A., Samph, T. & Case, S. (1978) *Short-term outcomes achieved in interviewing medical students*. Internal Report, National Board of Medical Examiners.

Van der Vleuten, C., Van Luyk, S., & Swanson, D. (1988) Reliability (generalizability) of the Maastricht Skills Test. *Proceedings of the Twenty-seventh Annual Conference of Research in Medical Education*, 228-233.

Van der Vleuten, C., Van Luyk, S., & Beckers, A. (1989) A written test as an alternative to performance testing. *Medical Education, 23*, 97-107.

Van der Vleuten, C., Van Luyk, S. Ballegooijen, A. & Swanson (in press) Training and experience of medical examiners. *Medical Education*.

Van Luyk, S., Van der Vleuten, C. & Peet, D. (1986) The assessment of clinical and technical skills at the medical school of Maastricht. In: Hart, I., Harden, R. & Walton, H. (Eds.) *Newer Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Vu, N., Steward, D. & Marcy, M. (1987) An assessment of the consistency and accuracy of standardized patients' simulations. *Journal of Medical Education, 62*, 1000-1002.

Williams, R. & Barrows, H. (1987) Performance-based assessment of clinical competence using clinical encounter multiple stations. In: Hart, I. & Harden, R. (Eds.) *Further Developments in Assessing Clinical Competence*. Montreal: Heal Publications.

Williams, R., Barrows, H., Vu, N., Verhulst, S., Colliver, J., Marcy, M. & Steward, D. (1987) Direct, standardized assessment of clinical competence. *Medical Education, 21*, 482-489.

Williams, R. & Colliver, J. (1987) A study of errors attributable to use of NBME Part I scores for residency selection purposes. *Proceedings of the Twenty-Sixth Annual Conference on Research in Medical Education*, 43-47.

Appendix 1: Summary of studies included in the review.

Institution	Reference(s)	Examinees	Station Format(s)
University of Adelaide	Newble & Swanson, 1988 Swanson & Norcini, in press	Senior Students	5-10 minutes for Physical, Patient Education, or Procedure
College of Family Physicians of Canada (CFPC)	Rainsberry et al., 1987a Grava-Gubins et al., 1987	Family Physician Candidates	15 minutes for History
Educational Commission for Foreign Medical Graduates (ECFMG)	Conn, 1986 Cody, 1988 Conn & Cody, under ed. review	Foreign Medical Graduates & U.S. Graduates	15-30 minutes for Physical, and 10-20 minutes for Written Followup[1]
University of Limburg	Van der Vleuten et al., 1988	Junior and Senior Students	10-20 minutes for History, or Physical or Procedure
University of Massachusetts (UMass)			
UMass Data Set 1	Stillman et al., 1986a Stillman et al., 1986b	1st, 2nd & 3rd Year Medicine Residents	30 minutes for History and Physical, and 10 minutes for Written Followup[1]
UMass Data Set 2	Stillman et al., 1987	Senior Students	10 minutes for History, and 5 minutes for Written Followup[1]
UMass Data Set 3	Stillman et al., under ed. review Swanson & Norcini, in press	Senior Students	15 minutes for History, and 5 minutes for Written Followup[1]
National Board of Medical Examiners (NBME)	Templeton et al., 1978	Senior Students	20 minutes for History and Initial Management

Appendix 1 continued on next page

Appendix 1 continued

Southern Illinois University (SIU)			
SIU Data Set 1	Colliver et al., 1989 Williams et al., 1987 Barrows et al., 1987 Williams & Barrows, 1987 Dawson-Saunders et al., 1987	Senior Students	15 minutes for History and Physical, and 15 minutes for Written Followup[2]
SIU Data Set 2	Klass et al., 1987 Tamblyn et al., 1988	Senior Students	15 minutes for History and Physical, and 15 minutes for Written Followup[2]
University of Texas Medical Branch at Galveston (UTMB)			
UTMB Data Set 1	Petrusa et al., 1984 Petrusa et al., 1986 Petrusa et al., 1987a	Junior Students	5 minutes for History or Physical, and 5 minutes for Written Followup[2]
UTMB Data Set 2	Petrusa et al., 1987b	1st and 2nd Year Medicine Residents	5 minutes for History or Physical, and 5 minutes for Written Followup[2]
UTMB Data Set 3	Petrusa, 1988	Junior Students	5 minutes for History or Physical, and 5 minutes for Written Followup[2]
University of Toronto	Cohen et al., 1987 Cohen et al., 1988	Foreign Medical Graduates	5-10 minutes for History or Physical, and 5 minutes for Written Followup[1]

[1]Written followup not included in testing time or calculation of scores in later tables
[2]Written followup included in testing time and calculation of scores in later tables

Samenvatting

Probleem-gestuurd leren is een succesvolle nieuwe onderwijsvorm gebleken, geïntroduceerd in vele instellingen voor hoger onderwijs. Probleem-gestuurd leren kan echter alleen succesvol zijn, als de wijze waarop studieprestaties van studenten geëvalueerd worden aansluit bij de principes van de onderwijsmethodiek. Bij de start van de Maastrichtse medische faculteit in 1974 was nog nauwelijks vastgesteld op welke wijze studieprestaties in dit nieuwe systeem geëvalueerd zouden worden. Begonnen werd met een tamelijk conventionele methode, waarbij iedere onderwijsleerperiode werd afgesloten met een toets. Al snel werd echter duidelijk dat deze aanpak op gespannen voet stond met de uitgangspunten van het onderwijssysteem, en dat naar alternatieve toetsprocedures moest worden gezocht. In 1982 startte de Faculteit der Geneeskunde een interdisciplinair project, dat tot doel had het evaluatiesysteem voor studieprestaties verder te ontwikkelen, aangepast aan de eisen van probleem-gestuurd leren. Dit proefschrift is binnen deze context tot stand gekomen.

Het proefschrift begint met een algemene beschrijving van het toetssysteem. Daarna volgen een viertal empirische studies over één onderdeel ervan: de toetsing van praktische vaardigheden door middel van zogenaamde vaardigheidstoetsen. Vaardigheidstoetsen bestaan uit een aantal 'stations', waarin studenten de opdracht krijgen praktisch medische verrichtingen daadwerkelijk uit te voeren. Deze vaardigheden worden beoordeeld door observatoren aan de hand van gedetailleerde criterialijsten. Toetsscores worden op basis van de ingevulde criterialijsten berekend. Hoewel nog weinig bekend was over de meeteigenschappen van toetsen voor praktische vaardigheden, zijn zij inmiddels in vele medische opleidingen geïntroduceerd. Pas in de laatste jaren zijn de eerste psychometrische studies verschenen. Door het geven van een overzicht en door bewerking van de resultaten van deze studies, beschrijft het laatste deel van dit proefschrift de huidige stand van zaken met betrekking tot de meeteigenschappen van observatietoetsen.

Hoofdstuk 1 beschrijft het huidige systeem voor evaluatie van studieresultaten dat bij de Maastrichtse medische faculteit wordt gebruikt. Vier zogenaamde competentie-domeinen worden in dit systeem onderscheiden: kennis, vaardigheden, probleemoplossen en attitudes. Voor elk van deze competenties volgt een beschrijving van formele en informele evaluatieprocedures, de ervaringen met deze procedures, de keuzes die werden gemaakt, en enkele empirische resultaten over betrouwbaarheid en validiteit. Voor de competentie-domeinen kennis en vaardigheden bestaat het huidige evaluatiesysteem uit een uitgebreid stelsel van bloktoetsen, voortgangstoetsen en vaardigheidstoetsen. De evaluatie van het competentiedomein probleemoplossen is beperkt tot een beoordeling van "medisch denken en handelen" in de stages en enkele informele evaluatieprocedures. Dezelfde situatie geldt voor de beoordeling van attitudes.

Benadrukt wordt dat een geïntegreerd evaluatiesysteem méér is dan de som van de gebruikte instrumenten en de hieruit resulterende test-scores. Daarom worden eveneens enkele algemene kenmerken van het evaluatiesysteem besproken. Deze kenmerken hebben betrekking op procedurele maatregelen voor kwaliteitscontrole van aangeleverd toetsmateriaal, studiebegeleiding als een integraal onderdeel van het examensysteem, en enkele organisatorische aspecten.

De conclusie is dat het huidige programma voor evaluatie van studieprestaties in een aantal opzichten de oorspronkelijke doelstellingen en uitgangspunten realiseert, maar dat op andere punten verbetering én uitbreiding nodig is. Met name zal aandacht besteed moeten worden aan nieuwe instrumenten voor het evalueren van probleemoplossen en, in een later stadium, aan evaluatie van attitudes. In verband met de wetenschappelijke stand van zaken op deze terreinen is het echter duidelijk dat op korte termijn weinig resultaten te verwachten zijn.

Eveneens wordt geconcludeerd dat het bestaande evaluatiesysteem, met grootschalige toetsprocedures en expliciet ingebouwde momenten voor kwaliteitscontrole, slechts tot stand kon komen door de centrale aanpak. Bovendien wordt gesteld dat deze aanpak een rationele, wetenschappelijke benadering van evaluatie van studieprestaties bevordert. Dat komt niet alleen door de mogelijkheden tot het uitvoeren van eigen onderzoeksprogramma's, maar ook door de 'transparantie' van het systeem, voor eenieder zichtbaar en open voor kritiek.

In **Hoofdstuk 2** wordt een studie beschreven over de betrouwbaarheid van de met vaardigheidstoetsen verkregen toetsresultaten. De toetsresultaten van alle zes de jaargroepen in de academische jaren 1984/85, 1985/86, en 1986/87 werden hiervoor geanalyseerd. Gebruik makend van de generaliseerbaarheidstheorie vonden berekeningen plaats ten aanzien van interbeoordelaarsbetrouwbaarheid en reproduceerheid van totale toetsscores.

De resultaten toonden aan dat de interbeoordelaarsbetrouwbaarheid voor alle jaren een adequaat niveau bereikte. Echter, de inconsistente prestaties van studenten van station tot station bleken een grote bron van onbetrouwbaarheid, ongeacht de jaargroep en de inhoud van de toets. Relatief lange toetstijden worden daardoor vereist voor het bereiken van een adequate betrouwbaarheid. Afhankelijk van de gekozen test-interpretatie lijkt een minimum toetslengte van drie tot vijf uur noodzakelijk. Suggesties voor verbetering van de betrouwbaarheid worden besproken.

Hoofdstuk 3 gaat specifiek in op de accuratesse waarmee beoordelaars of observatoren de prestaties van studenten waarderen. In eerdere studies was gebleken dat training van examinatoren slechts een marginale verbetering van de accuratesse opleverde. In dit hoofdstuk wordt verslag gedaan van een experiment waarin de invloed van zowel training als deskundigheid van beoordelaars werd onderzocht. Artsen, medische studenten en leken werden random toegewezen aan een groep die wel en een groep die geen training onderging. Gebruik makend van criterialijsten scoorden zij een tweetal op videoband geregistreerde stations. Vervolgens werd hun accuratesse bepaald.

De resultaten toonden dat het nut en de efficiëntie van de beoordelaarstraining varieerde met het deskundigheidsniveau: marginale verbetering in accuratesse werd geboekt bij de artsen, een groter effect was zichtbaar bij de medische studenten, en het grootste effect werd behaald bij de leken. De accuratesse van de getrainde groep leken benaderde die van de ongetrainde artsen, terwijl de getrainde medische studenten even accuraat bleken als de getrainde artsen. Voor de artsen en medische studenten bleek dat training eveneens effect had op de kwaliteit van de beoordeling: het aantal fouten met bepaalde structurele fouten nam af.

Geconcludeerd wordt dat de effectiviteit van beoordelaarstraining varieert met de expertise van de beoordelaar, en dat getrainde leken kunnen worden gebruikt als beoordelaars in observatietoetsen.

In **Hoofdstuk 4** wordt verslag gedaan van een studie naar de mogelijkheid van schriftelijke meting van praktische vaardigheden. Een kennistoets over vaardigheden (KOV) werd ontwikkeld en afgenomen bij 380 personen van verschillend opleidingsniveau, waaronder startende medische studenten en recent afgestudeerde artsen (uit heel Nederland).

Sterk convergerende validiteit bleek bij het vergelijken van KOV-scores met scores op de vaardigheidstoets. Discriminerende validiteit door vergelijking met scores op een algemeen medische kennistoets kon echter niet worden aangetoond. Ook de identificatie van delen van de KOV, die meer gedragsmatige dan wel cognitieve aspecten zouden meten, was niet succesvol. Dit werd verklaard door de onderlinge afhankelijkheid van de gemeten constructen. De construct-validiteit werd ondersteund door de bevindingen dat de KOV in staat was te discrimineren tussen groepen van verschillend kennisniveau en doordat verwachte veranderingen in antwoordpatronen konden worden aangetoond.

De conclusie is dat de KOV, met name in de hogere jaargroepen, een valide instrument is voor het voorspellen van praktische vaardigheden en dat soortgelijke instrumenten goed gebruikt zouden kunnen worden voor het verkrijgen van additionele informatie over praktische vaardigheden. De eenvoud van constructie en de efficiëntie van de toetsvorm, maken de KOV ook een geschikt alternatief instrument voor de meting van praktische vaardigheden ten behoeve van onderzoeksdoeleinden.

Uit dit onderzoek blijkt eveneens dat naarmate het niveau stijgt de gemeten concepten onderling een sterkere samenhang vertonen. Deels steunt dit de theorie dat medische competentie kan worden verklaard op grond van één enkele algemene factor, vergelijkbaar met de g-factor in intelligentie-onderzoek. Deze factor is afwezig bij de lagere niveau's. Naarmate het onderwijsleerproces vordert 'integreren' echter de afzonderlijke competenties.

Hoofdstuk 5 beschrijft een validiteitsstudie van de vaardigheidstoets, waarin werd nagegaan in hoeverre de sterk analytische beoordeling door middel van criterialijsten afwijkt van algemene indrukken van observatoren. Scores op criterialijsten werden daartoe vergeleken met een globaal oordeel van de observator. Ter verklaring van mogelijke discrepanties werd bovendien een aantal meer specifieke oordelen gevraagd, betrekking hebbend op aspecten die in meer of mindere mate door de bestaande criterialijsten werden gedekt. Deze

aspecten bestonden uit: kwaliteitsaspecten van de techniek van de verrichte handelingen, probleemoplossingsvaardigheden, persoonlijkheidsfactoren en attitude van de student. Door vergelijking van deze specifieke oordelen met de scores op criterialijsten werd een poging ondernomen de construct-validiteit van de vaardigheidstoets nader te onderbouwen.

Uit de resultaten bleek dat er geen sprake was van grote verschillen tussen uitkomsten van criterialijsten en het algemene oordeel van observatoren. In strijd met de verwachting bleek dat de aanwezige discrepanties niet zonder meer te verklaren waren door de aspecten die in mindere mate door de criterialijsten worden gedekt. Verschillen bleken eveneens op te treden bij die aspecten waarvan verondersteld mag worden dat criterialijsten een objectievere meting opleveren (zoals de technische aspecten).

De praktijkomstandigheden in deze studie maakten het noodzakelijk dat de globale beoordelingen gegeven werden na invulling van de criterialijsten, waardoor van een zekere beïnvloeding sprake kan zijn geweest. In enkele controle-experimenten bleek deze invloed echter verwaarloosbaar klein.

In **Hoofdstuk 6** wordt de huidige stand van zaken in de wetenschap geschetst met betrekking tot observatietoetsen. Hoewel bij observatietoetsen een grote verscheidenheid aan methoden mogelijk is, wordt meestal gebruik gemaakt van simulatiepatiënten (SP). Dit zijn niet medisch geschoolde personen, die getraind zijn om op betrouwbare wijze een patiënten-rol te spelen (in de meest brede betekenis: zij kunnen ook gewoon proefpersoon zijn voor een lichamelijk onderzoek). In het laatste decennium zijn de observatietoetsen die gebruik maken van deze methode zeer populair geworden en zijn vele instellingen tot introductie van dergelijke toetsen overgegaan. Niettemin is nog weinig bekend over de meeteigenschappen van deze instrumenten.

De laatste jaren zijn een aantal studies gedaan naar de psychometrische eigenschappen van observatietoetsen. Dit hoofdstuk bespreekt en analyseert de resultaten hiervan. Het overzicht wordt beperkt tot observatietoetsen waarbij van de SP-methode gebruik wordt gemaakt.

Op consistente wijze tonen de betrouwbaarheidsanalyses van de verschillende studies aan dat de grootste foutenbron wordt veroorzaakt door de inconsistente prestaties van studenten over verschillende stations, in de literatuur over probleem-oplossen ook wel aangeduid als het casus-specificiteitsprobleem. Het heeft tot gevolg dat grote aantallen stations in toetsen moeten worden opgenomen om een adequate betrouwbaarheid te bewerkstelligen. Dit leidt noodzakelijkerwijs tot lange toetstijden. Andere foutenbronnen, zoals beoordelaarsverschillen en verschillen tussen simulatiepatiënten die eenzelfde rol spelen, hebben een veel geringer effect op de precisie van de test-scores.

De resultaten van conventionele validiteitsstudies, met name studies van groepsverschillen en correlaties met andere metingen, blijken in het algemeen positief, maar niet bijzonder informatief. Voorgesteld wordt dat toekomstig validiteitsonderzoek zich zou moeten richten op:

1. Studie van de invloed van stationseigenschappen, zoals inhoud, tijdsduur en instructies, op de prestaties van studenten.
2. Studie van de wijze waarop prestaties van studenten worden vertaald in stations- en toetsscores.

Aanbevelingen voor de verbetering van SP-gebaseerde observatietoetsen worden besproken. Deze omvatten:

1. Beperking van de inhoud tot de meer handelingsgerichte vaardigheden, waarbij eventueel separate schriftelijke toetsen kunnen worden gebruikt voor de meer cognitief georiënteerde vaardigheden
2. Interpretatie van test-scores volgens een domein-georiënteerd perspectief, eventueel gebruik makend van sequentiële testprocedures.
3. Ontwikkeling van procedures voor zak/slaag beslissingen.

Het gebruik van generaliseerbaarheidstheorie voor onderzoek en rapportage wordt eveneens geadviseerd.

Summary

Problem-based learning is now acknowledged to be a successful educational method, and it has been adopted in many institutions in higher education. However, for problem-based learning to be successful, the system used for assessment of student achievement must be consistent with the educational principles of problem-based learning.

When students first matriculated into the Maastricht medical school in 1974, the school's assessment philosophy and system were not well developed. Initially, the school used a traditional approach: tests were given at the end of each educational unit. However, it shortly became apparent that this strategy conflicted with the design and intentions of the educational program. In 1982, the medical school adopted a more centralized system of assessment, and a multi-disciplinary group of faculty was given the task of designing, implementing, and maintaining an assessment system specifically adapted to the problem-based learning method. The work described in this dissertation was conducted within this context.

The first chapter provides an overview of the assessment system used by the Maastricht medical school. The following four chapters describe psychometric studies of one component of the assessment system, the Skills Test. The Skills Test is designed to measure examinees' ability to perform a variety of "hands-on" technical and clinical tasks, with faculty-observers scoring examinee performance on detailed behavioral checklists. Despite a lack of information about the measurement characteristics of such exams, performance-based assessments similar to the Skills Test have become very popular in the past decade and have been introduced in many medical schools and other institutions providing education in the health professions. The final chapter of the dissertation reviews other studies of the measurement characteristics of performance-based tests and summarizes the current state of the art from a psychometric perspective.

Chapter 1 describes the current assessment system at the Maastricht medical school. Four competencies are measured: knowledge, skills, problem-solving, and attitudes. For each of these, the formal and informal components of the assessment system are described, the rationale underlying the design is provided, practical experience in conducting the assessment is summarized, and results of reliability and validity studies are outlined.

The current system includes an elaborate program of formal assessments of knowledge and skills, consisting of Block Tests, Progress Tests, and Skills Tests. With the exception of Clinical Ratings, assessment of problem-solving and attitudes is still informal.

Design of an assessment system must include more than independent construction and use of testing instruments. The overall goals and organization of

the assessment system are also discussed, along with procedures for quality control of test materials and interrelationships with student counseling and promotion.

It is concluded that the current assessment system is consistent with the principles of problem-based learning and meets many of the school's needs, but that further refinement of the system is desirable. Methods for assessment of problem-solving and attitudes are still needed, though results cannot be expected in the short term, because of the current state of the art in measuring these competencies. A centralized approach to assessment permits use of more elaborate testing methods, without loss of control over quality, a likely problem if a traditional, decentralized approach were used. The centralized approach also allows systematic, scientific development of assessment methods that are open to public scrutiny.

In **chapter 2**, a study of the reproducibility of Skills Test scores is reported[1]. The study included examinee scores from all classes in the 1984/85, 1985/86, and 1986/87 academic years. Using methods derived from generalizability theory, analyses investigated inter-rater reliability of station scores and overall reproducibility of total test scores.

Inter-rater reliability was adequate across all classes. Variation in the quality of examinee performance from station to station proved to be a much larger source of measurement error. Decision studies indicated that, depending upon the perspective adopted for interpretation of scores, a minimum of three to five hours of testing time is required to obtain reasonably reproducible scores, regardless of year of training and despite differences in test content. Some strategies for increasing reliability are discussed.

Chapter 3 focusses in more detail on the accuracy of scores based upon examiner ratings. In previous studies using physician-subjects, examiner training resulted in marginal improvement in the accuracy of examiner judgments. This study systematically investigated the impact of training and examiner background on accuracy of scores. Experienced faculty, medical students, and lay subjects were randomly assigned to either Training or No-Training groups. Using detailed behavioral checklists, they subsequently scored videotapes of examinees working though two clinical cases, and the accuracy of their judgments was appraised.

Results indicated that the need for and effectiveness of training varied across groups: it was least needed and least effective for members of the faculty group, more needed and effective for medical students, and most needed and effective for the lay group. The accuracy of the lay group after training approached the accuracy of untrained faculty. Trained medical students were as accurate as trained faculty. For faculty and medical students, training also

---

[1]Chapter 2 is written in the Dutch language. For an English text on the reproducibility of Skills Test scores referral is made to Van der Vleuten, C.P.M., Luyk, S.J. van & Swanson, D.B. (1988) Reliability (generalizability) of the Maastricht Skills Test. *Proceedings of the Twenty-seventh Annual Conference on Research in Medical Education (RIME)*, Chicago, USA.

influenced the nature of the errors made by reducing the number of errors of commission. It was concluded that training varies in effectiveness as a function of medical background and that trained lay persons can serve as examiners in performance-based tests without serious loss of accuracy.

In **chapter 4**, a study is reported which investigated the use of a written examinations as an alternative to performance-based testing. A knowledge test of skills (KTS) was developed and administered to 380 subjects at various educational levels, ranging from first year medical students to recently graduated physicians.

By comparing scores on the KTS with scores on performance tests, strong convergent validity was demonstrated. However, no discriminant validity was obtained when scores on the KTS were compared with performance on a test of general medical knowledge: these scores were also highly correlated. In addition, identification of subtests discriminating between behavioral and cognitive aspects was not successful. The KTS was able to detect differences in ability level, and the pattern of changes in response patterns over items provided additional evidence of construct validity. It was concluded that the KTS accurately predicted scores on performance-based tests and could be used as a supplement to performance-based testing. Relative ease in test construction and efficiency in use of testing time make the KTS an attractive substitute for performance-based tests in research efforts.

The study also showed that stronger relationships existed at more advanced levels of training, between the constructs measured with the different instruments, supporting a general factor theory of competence. However, it appeared that this general factor was originally non-existent in freshmen and that these competencies developed as the educational process continued.

To establish concurrent validity of the Skills Test, **chapter 5** reports a study in which examinee scores based upon checklists from Skills Test stations (analytic judgment procedure) were compared with the general impressions of expert observers recorded on a rating scale (global judgment procedure). In addition, observer ratings of problem solving, attitudes, personality, and technique were compared with Skills Test outcomes, allowing a more sensitive examination of the construct validity of the latter.

Results showed no gross differences between the global and analytic judgment methods, in terms of highly and poorly rated performance. Somewhat surprisingly, however, discrepancies were observed between checklist-based scores and ratings of components in areas in which the Skills should be most effective (i.e. technique).

Since global ratings were gathered during regular Skills Test administrations by regular examiners after they had scored with checklists, completion of global ratings may have been biased. In control experiments this alternative explanation was falsified.

Chapter 6 reviews the current state of the performance-based testing art, focussing on standardized patients (SPs) -- non-physicians trained to reproducibly play the role of a patient for assessment purposes. Most performance-based

tests use SPs at several stations to assess clinical skills. In the past few years, a number of studies have investigated the psychometric characteristics of SP-based tests; this chapter provides a comprehensive review of this work.

Across studies, reliability analyses consistently indicated that the major source of measurement error is variation in examinee performance from station to station, also known as "case-specificity" in the medical problem solving literature. As a consequence, it is necessary to include large numbers of stations in order to obtain a stable, reproducible assessment of examinee skills. Somewhat surprisingly, disagreements among raters observing examinee performance and differences between SPs playing the same patient role have much less effect on the precision of scores.

Results of conventional validity analyses (e.g., studies of group differences; correlations with other measures) were generally favorable, though not particularly informative. It is argued that future validity research should include:

1. Study of the impact of station format, timing, and instructions on performance.
2. Study of procedures used to translate examinee behavior into station and test scores.

Several recommendations are offered for improvement of SP-based tests. These include:

1. Focussing on assessment of history-taking, physical examination and communication skills, with separately administered written tests used to measure diagnostic and management skills.
2. Adoption of a mastery-testing framework for score interpretation.
3. Further development of standard setting procedures.

Use of generalizability theory in analyzing and reporting future psychometric research is also suggested.

Curriculum vitae

C. van der Vleuten werd op 6 juni 1956 geboren te Eindhoven. Hij voltooide zijn Atheneum-A opleiding in 1975 aan het Eckart College te Eindhoven. In datzelfde jaar begon hij zijn studie psychologie aan de Katholieke Universiteit Brabant. In 1982 werd het doctoraal examen behaald (cum laude), met als specialisatie Persoonlijkheidsleer en Psychodiagnostiek. Sedert 1982 is hij als universitair docent verbonden aan de vakgroep Onderwijsontwikkeling en Onderwijsresearch van de Rijksuniversiteit Limburg. Eerst als lid en later als projectleider is hij actief betrokken bij het Project Evaluatie van Studieresultaten van de Faculteit der Geneeskunde.